Identification of Vegetation State-and-transition Domains in California's Hardwood Rangelands.

By

Marc P. Vayssières
and Richard E. Plant

Agronomy and Range Science
Hunt Hall, One Shields Ave
University of California
Davis, California 95616

for

Fire and Resource Assessment Program
California Department of Forestry and Fire Protection
1920 20th Street
Sacramento, California 95814

May 1998

Table of content

List of tables

List of Figures

Introduction.

Bioregional planning, assessment and monitoring of natural resources require predicting spatial patterns of vegetation at the landscape level. Understanding vegetation factors and processes is necessary to predict future patterns of vegetation in landscapes. There is a renewed interest in implementing models of vegetation dynamics to assess the effect of human activities on ecosystems and help manage landscapes. In range science, the traditional Clementsian approach has proven inadequate and state-and-transition models have recently been proposed as an alternative. The development of such models is a promising way to synthesize our understanding of California ŝ hardwood rangelands.

The state-and-transition framework.

The traditional succession model, based on Clements 'ideas of plant succession (Clements, 1916, Weaver and Clements, 1938), was first proposed for rangelands by Sampson (1917) who suggested that vegetation changes under grazing management were equivalent to the changes occurring during secondary succession. Later, Dyksterhuis (1949, 1958) formalized the range succession model and promoted what remains the most influential method to evaluate range condition. The traditional model is based on the following assumptions: a particular rangeland site has a unique climax state; succession toward the climax is a linear, continuous, monotonic and reversible process; grazing pressure produces continuous changes directly opposite to the successional tendency; and variations in climate have effects similar to those of grazing. However, documented cases where these assumptions do not hold have accumulated over the years.

By the end of the 1980 ŝ, the many problems of the climax-based approach to rangeland dynamics had made a change of paradigm unavoidable (Smith, 1989; Wilson, 1989; Friedel, 1991; Laycock, 1991). During the same period ecologists had abandoned the notions of deterministic succession and single equilibrium communities as too simplistic. They have replaced them by the concepts of multiple successional pathways, thresholds, alternative stable states, and discontinuous transitions affected by plant strategies, initial conditions and stochastic events (Connell and Slatyer, 1977, Holling 1973; May 1977; Nobel and Slatyer 1980; Noy-Meir 1982; Godron and Forman 1983; Taush et al. 1993).

In 1989, Westoby et al. proposed to describe rangeland dynamics at a particular site by a set of discrete persisting *states* of the vegetation and a set of possible *transitions* between these states. The transitions may be caused by natural disturbances (e.g., weather, fire, herbivory) or by management actions (e.g., grazing, burning, wood harvest, elimination or introduction of plant species, fertilization). The states describe alternative and stable vegetation types but, because they are a generalization, they may include a certain amount of variation in time and space. The State-and-transition (S&T) concept is not a new all-encompassing theory, but rather a framework to summarize knowledge of vegetation dynamics without distorting it. Indeed, dynamics consistent with the traditional linear succession model could be described by a S&T model.

The S&T formulation has many advantages. It can accommodate practical experience as well as experimental results, qualitative and quantitative knowledge. S&T models can be used to organize incomplete understanding and then grow as new knowledge becomes available. They structure information in a way that can help managers focus on opportunities to direct change toward favorable states and to avoid transitions that are irreversible or too costly to reverse in term of management intensity, labor and inputs. In sum, they provide a framework for a pro-active land management instead of the fatalistic view promoted by the old successional model.

Typically, S&T models have been implemented through simple printed flowcharts complemented by catalogs of state and transitions. For instance, George et al. (1992) designed a S&T model of the dynamics of oak savanna, shrubland or grassland for a foothill range site in Yuba county, California. And Huntsinger and Bartolome (1992) have described a more generic model for oak woodlands and savannas with a potential shrub understory in California and Spain. However S&T models can also be implemented on a computer using expert system methodologies. Computerized implementation and linkage to a geographic information system allow for testing of the model through spatially explicit simulations (Plant et al., in press). Spatially explicit models based on S&T models could be very useful to rank management alternatives at the landscape level.

Building state-and-transition models.

Although S&T models have been the subject of much academic discussion, they have as yet had little practical application. Most of the models that have been

proposed to date do not go much beyond proof-of-concept examples. There is definitely a need for methodologies to help build more operational models.

As a first step, S&T models should be used as a framework to organize current understanding about vegetation dynamics. Much has been published, for instance, about the ecology, disturbances and management of California oak woodlands, and bibliographies have been compiled (e.g., Griffin et al. 1987). However, most published papers report reductionist research work, and integrating this knowledge in a way that can support management decisions is a task too complex for any single person. Bellamy and Brown (1994) proposed to use an iterative process for summarizing knowledge into states and transitions, including literature searches, interviews and workshops with experts, and evaluations of the knowledge base by land managers and end-users.

There are various ways to supplement such a qualitative process by quantitative analyses depending on the type of data on hand. When temporal vegetation data are available, classification and ordination methods can help determine some of the alternate states of a system (Friedel, 1991; Stafford-Smith and Pickup, 1993). Experimental data have also a role to play in refining S&T models. Conventional experiments can help identify thresholds between states and the mechanisms involved in transitions (Jones, 1992; Jameson, 1991). Allen-Diaz and Bartolome (in press) combined these two approaches by classifying and quantitatively identifying states and transitions observed over a period of 20 years following experimental range improvement treatments. When one-time vegetation data is complemented by disturbance history and physical site characteristics, classification and ordination methods can also help delineate states and identify transitions for a particular site (Bork et al., 1997). Unfortunately, it is often difficult to reconstruct the disturbance and management history corresponding to changes in vegetation observed *a posteriori*. When historical data are available, it is sometimes possible to revisit sites where the vegetation was surveyed years before. For instance, Holzman and Allen-Diaz (1992) relocated and surveyed oak woodland vegetation plots taken in the 1930 s. Plant et al. (in press) had management history and used a sequence of 5 aerial photographs to characterize vegetation changes at a hardwood range site between 1952 and 1993. However, these two studies showed that even 40- and 60-year intervals do not reveal much change in these long-lived communities. Although the ideal would be to have long-time monitoring data (including information on vegetation, natural

disturbances and management actions) to develop S&T models, such data are just too scarce.

One solution is to substitute an understanding of vegetation variation in space for knowledge of variations in time to delineate groups of vegetation states susceptible to be linked by transitions. It is reasonable to assume that, at the regional or landscape level, vegetation states present at one point in time are representative of the variety of states that could be found over time. Indeed, the vegetation mosaic observed in landscapes is both the result of temporal (i.e., successional) processes occurring asynchronously in space (Watt 1947), and a response to various environmental gradients resulting from spatial processes (Whittaker 1953). Inspired by Jenny (1941), Major (1951) proposed what amounts to a synthesis of these views, arguing that vegetation is a function of regional climate, soil parent material, topography or relief, organisms (flora, fauna and humans) and time. Later, Jenny (1961) extended this approach to the formation of ecosystems. Using such a factorial approach as a framework can help us in designing S&T models, particularly at the landscape level.

Jenny (1961) classed vegetation/ecosystem state factors into two groups: (1) the "initial state" factors, parent material and topography, that are not time dependent on the scale we are concerned with; (2) the "external flux" factors, regional climate and organisms (flora, fauna, humans), that may be functions of the time factor. Jenny's classification of factors overlaps nicely with the amount of control that management can exert on these factors. At one end of the spectrum, parent material and topography cannot be modified, but they do not vary over time either: they can be seen as constants which set the stage for the interactions of other factors. As these two factors vary over a landscape, their patterns partly account for the vegetation mosaic. Climate is more variable over time, especially in arid and semi arid regions. At the other end of the spectrum, the presence of organisms and their impact are the most variable factors and, at the same time, the most amenable to management influence. The contribution of any state factor to changes in vegetation in a chosen landscape becomes negligibly small either if the factor is almost constant in the area or if its influence is small (Jenny 1961). Spatial analyses of the correlations of vegetation data with climate, parent material and topographic position can help delimit regions (or landscape positions) where these factors are sufficiently constant in their influence on vegetation. Within these regions, the observed variations in vegetation should be the result of disturbance

and management influence. In other words, if we can identify domains where the structural "limiting" factors are relatively constant in their influence, then the states found in these domains are potentially linked by transitions.

Another useful way to think about separating the influence of spatial position from the influence of disturbances in constructing S&T models is in terms of the "assembly rules" and "response rules" of Keddy (1989, 1992). Transitions between states are equivalent to the response rules, as they depend on the response of plant species to disturbances, to fluctuations in the environment and to competition. Any understanding of the assembly rules (the filtering that the environment exerts on the regional pool of species) can greatly reduce the set of possible outcomes and thus help us define potential vegetation states. Thus, the identification of states and transitions can clearly benefit from a functional approach to plant ecology. Eventually, functional classifications of vegetation will provide the basis for more detailed and usable S&T models. Because it is based on presence/absence and relative abundance of individual plant species, conventional vegetation classification work often results in vegetation types that are at the same time too specific and too general to serve as potential vegetation states. Classifying vegetation on the basis of functional groups of plants is more appropriate to define states and transitions. The functional attributes reflecting the mechanisms involved in transitions include life history (annual, short-lived, long-lived perennial), palatability, resistance to grazing and browsing, resistance and adaptation to fire (e.g., ability to resprout, fire germination), deciduousness, drought tolerance, nitrogen fixation, etc. Recent work aiming at the derivation of functional groups to simplify vegetation data in Mediterranean and semi-arid climates includes Friedel et al. (1988), Leishman and Westoby (1992) and Fernández-Alés et al. (1993). A functional approach is necessary to separate further the effects of management from those of other factors.

Our goal is to delimit domains within the California's oak woodlands for which consistent S&T models could be developed using expert workshops. Oak woodland vegetation types in California are diverse. The overstory may be dominated by one or several of six main *Quercus* species or their hybrids: *Q. douglasii, Q. Kelloggii, Q. agrifolia, Q. wislizenii, Q. lobata,* and *Q. englemanii* (blue, black, coast live, interior live, valley and Englemann oaks); and the shrub layer, if present, may include a combination of many species. Tree density ranges from sparse savannas to closed canopy forests. European settlement of California

has brought profound ecological changes to those woodlands. In particular to the grass layer where species from the Mediterranean basin have replaced the native understory of grasses and forbs, modifying the competitive environment of oak seedlings. Allen et al. (1991) have classified California's oak woodland using multivariate techniques (TWINSPAN, DECORANA) on data from several thousand survey plots. The resulting cover types have been used by Vayssières et al. (1993) to develop a spatially explicit model to forecast the response of California's oak woodlands to fire, grazing, browsing, and wood cutting. However, this classification proved unsatisfactory to describe vegetation dynamics and deriving functional groups has been difficult, at least from information currently available. Attempts to develop functional types based on longevity, fire resistance, ability to resprout after a fire, and response to grazing and browsing (George et al. 1993) were hampered by the difficulty of finding such knowledge in published sources. Thus, in this paper we are developing a simpler scheme based on combination of life forms into physiognomic groups.

We propose that on larger spatial scales, vegetation states of interest for the management of California's hardwood rangelands are somewhat independent of the presence or absence of particular species. Similar states though, may be reached through various transitions depending on species particularities and local environmental factors. Huntsinger and Bartolome (1992) found that vegetation states in the Spanish and Californian oak woodlands had similar appearance and function. However the ecological dynamics varied in some important ways, so that vegetation states are reached and are maintained in different fashion in each landscape.

In California's Mediterranean climate, water balance is the major environmental gradient determining species composition and vegetation physiognomy at a site. Water balance at the patch level is a complex function of latitude, distance from the coast, orographic effects, exposition, slope and topographic position, soil depth and texture, and opportunity to access a water table. There is some indication that the same basic vegetation states occur throughout the oak woodlands but that transitions leading to these states depend both on species specificities and local conditions. For instance: live oaks regenerate more easily than deciduous ones (Bartolome et al. 1987); coast live oak is extremely fire resistant while canyon live oak is quite fire sensitive (Plumb 1979); blue oak resprouts readily north of Madera County but very little in the

south; in southern Sierra counties, blue oak sapling recruitment increases with the amount of orographic precipitation (Standiford et al. 1991); oak canopy lessens forage production where precipitation exceeds 50 cm per year but favors it where it is less (McLaran and Bartolome 1989).

In this paper we reclassify the plots classified by Allen et al. into physiognomic groups. Then we relate these groups to climatic, topographic and soil factors to delineate abiotic domains containing particular combinations of physiognomic groups. Lastly, we show how this information can be used by a group of expert as the basis for the definition of a S&T model for a particular domain. Because species information is not determinant to identify potential vegetation states but is important to define the transitions, we have designed a new approach to cluster analysis to generalize from species-based cover types to physiognomic groups while retaining some of the species information. We call this new approach supervised conceptual clustering. This approach is then extended to the grouping of physiognomic groups into abiotic domains.

Methods.

The data.

Our data are based on survey plots from a historic dataset, the Vegetation Type Map (VTM) survey of California. This survey, conducted in the 1920's and 1930's, remains the most extensive systematic sampling effort of California's vegetation. Field crews gathered data from more than 8000 plots in the coastal ranges from San Francisco bay to the Mexican border and a large part of the Sierra Nevada. VTM data served as a basis for a statewide effort to classify and map vegetation (Jensen, 1947) and for Griffin and Critchfield's (1972) work on the distribution of forest tree species. Allen and coworkers entered 4288 VTM records (all the plots featuring a *Quercus* species) in a database and analyzed 2038 of them to construct a classification system for California's hardwood rangelands (Allen et al. 1989, 1991). They identified patterns in species distribution using TWINSPAN (two-way indicator species analysis; Hill, 1979a), with pseudo-species cut levels adjusted to accommodate both cover values in percent and tree basal area in square feet per acre. After initial analyses in four different geographical regions, plots with similar oak species dominance were reanalyzed together. The program DECORANA (detrended correspondence analysis; Hill, 1979b) was then used to examine relationships between types through available environmental gradients. Allen et al. went through a process of continuous feedback analysis using TWISPAN, DECORANA, ANOVA and regression techniques to increase cover type homogeneity and finalize their classification, including type descriptions and identification keys. The classification was also tested in the field and reviewed by a number of hardwood experts. The California's hardwood rangelands cover types –also known as Allen classes– comprise 57 subseries within 7 series (Table 1). Later, Evett (1994) added geographic coordinates and climatic data to the database and built models of the environmental niche for six oak species using the direct gradient analysis approach of Austin et al. (1990).

Table 1: California ś hardwood rangelands cover types (Allen et al. 1989, 1991).

#	Short Name	Subseries name	Plots
1	Qudu2-Qudo/Grass	Quercus dumosa-Quercus Douglasii / *Grass*	14
2	Qudu2/Grass	*Quercus dumosa* / Grass	16
3	Qudu2	Quercus dumosa	11
4	Mo-Pisa2/Grass	Mixed *Quercus* sp.-*Pinus Sabiniana* / Grass	65
5	Qudo-Pisa2/Grass	*Quercus Douglasii-Pinus Sabiniana* / Grass	98
6	Qudo-Pisa2/Cecu2-Cebe2	*Quercus Douglasii-Pinus Sabiniana / Ceanothus cuneatus-Cercocarpus betuloides*	23
7	Qudo/Cecu2/Grass	*Quercus Douglasii / Ceanothus cuneatus* / Grass	47
8	Qudo-Quwi/Grass	Quercus Douglasii-Quercus Wislizenii / *Grass*	70
9	Quwi-Qudo-Pisa2/Grass	*Quercus Wislizenii-Quercus Douglasii-Pinus Sabiniana* / Grass	44
10	Qudo-Pisa2/Arvi3/Grass	*Quercus Douglasii-Pinus Sabiniana / Arctostaphilos viscida* / Grass	60
11	Quwi-Arme3/Rhdi	*Quercus Wislizenii-Arbutus Menziesii / Rhus diversiloba*	23
12	Quwi/Erca6/Grass	*Quercus Wislizenii / Eriodictyon californicum* / Grass	36
13	Quwi-Pisa2/Arma3	*Quercus Wislizenii-Pinus Sabiniana / Arctostaphylos manzanita*	36
14	Quwi/Arvi3	*Quercus Wislizenii / Arctostaphilos viscida*	72
15	Quwi/Hear2	*Quercus Wislizenii / Heteromeles arbutifolia*	42
16	Quag/Adfa-Same4	*Quercus agrifolia / Adenostoma fasciculatum-Salvia mellifera*	20
17	Quag/Arca7/Grass	*Quercus agrifolia / Artemisia californica* / Grass	113
18	Quag	*Quercus agrifolia*	9
19	Quag-Arme3/Coco5-Ruvi2	*Quercus agrifolia-Arbutus Menziesii / Corylus cornuta-Rubus sp.*	23
20	Quag/Ruvi2/Ptaq	*Quercus agrifolia / Rubus sp. / Pteridium aquilinum*	18
21	Quag/Grass	*Quercus agrifolia* / Grass	13
22	Quke-Qulo/Grass	*Quercus Kelloggii-Quercus lobata* / Grass	19
23	Qulo/Grass	*Quercus lobata* / Grass	44
24	Qulo-Quag/Grass	*Quercus lobata-Quercus agrifolia* / Grass	37
25	Quag-Qulo/Rhdi	*Quercus agrifolia-Quercus lobata / Rhus diversiloba*	34
26	Mo-Qulo/Rhdi-Rhca2	*Mixed Quercus sp.-Quercus lobata / Rhus diversiloba-Rhamnus californica*	22
27	Mo/Grass	*Mixed Quercus sp.* / Grass	34
28	Qudo-Qulo/Grass	*Quercus Douglasii-Quercus lobata* / Grass	37
29	Mo-Aeca2/Grass	*Mixed Quercus sp.-California Buckeye* / Grass	29
30	Qudo-Quag/Grass	*Quercus Douglasii-Quercus agrifolia* / Grass	16
31	Qudo/Grass	*Quercus Douglasii* / Grass	291
32	Qudo/Uqudo/Grass	*Quercus Douglasii* / Understory *Quercus Douglasii* / Grass	61
33	Qudo/Hali	Quercus Douglasii / Haplopappus linearifolius	17
34	Qudo-Qulo-Quag/Grass	*Quercus Douglasii-Quercus lobata-Quercus agrifolia* / Grass	10

#	Short Name	Subseries name	Plots
35	Quke/Rhdi-Stofc/Brla2	*Quercus Kelloggii / Rhus diversiloba-Styrax officinalis / Brodiaea laxa*	19
36	Quke/Rhdi	*Quercus Kelloggii / Rhus diversiloba*	19
37	Quke-Arme3-Quag	*Quercus Kelloggii-Arbutus Menziesii-Quercus agrifolia*	23
38	Mo-Quag/Rhdi	*Mixed Quercus sp.-Quercus agrifolia / Rhus diversiloba*	42
39	Quke-Quag-Pico1/Hodi	*Quercus Kelloggii-Quercus agrifolia-Pinus contorta / Holodiscus discolor*	17
40	Mo-Quke/Grass	Mixed *Quercus sp.-Quercus Kelloggii* / Grass	36
41	Quke-Rhdi/Grass	*Quercus Kelloggii / Rhus diversiloba* / Grass	13
42	Quke/Grass	*Quercus Kelloggii* / Grass	15
43	Quag-Umca1/Hear2-Qudu2	*Quercus agrifolia-Umbellularia californica / Heteromeles arbutifolia-Quercus dumosa*	21
44	Mo/Rhdi-Bapi	Mixed *Quercus sp. / Rhus diversiloba-Baccharis pilularis*	22
45	Quag/Rhca2-Hear2	*Quercus agrifolia / Rhamnus californica-Heteromeles arbutifolia*	33
46	Quag/Hear2-Rhdi	*Quercus agrifolia / Heteromeles arbutifolia-Rhus diversiloba*	48
47	Quag/Hear2/Grass	*Quercus agrifolia / Heteromeles arbutifolia* / Grass	13
48	Quag/Rhdi	*Quercus agrifolia / Rhus diversiloba*	47
49	Quag/Rhdi/Grass	*Quercus agrifolia / Rhus diversiloba* / Grass	40
50	Quag/Hodi-Syri	Quercus agrifolia / Holodiscus discolor-Symphoricarpos rivularis	20
51	Quag-Acma/Rhca2-Hodi	*Quercus agrifolia-Acer macrophyllum / Rhamnus californica-Holodiscus discolor*	11
52	Mo-Quwi-Pisa2	Mixed *Quercus sp.-Quercus Wislizenii-Pinus Sabiniana*	9
53	Quke-Quch2/Rhdi	*Quercus Kelloggii-Quercus chrysolepis / Rhus diversiloba*	19
54	Quke/Arpa9	Quercus Kelloggii / Arctostaphylos patula	19
55	Quke/Cein3-Rhdi/Ptaq	*Quercus Kelloggii / Ceanothus integerrimus-Rhus diversiloba / Pteridium aquilinum*	11
56	Quke/Cein3	*Quercus Kelloggii / Ceanothus integerrimus*	13
57	Quch2-Quke	*Quercus chrysolepis-Quercus Kelloggii*	24

Note: Scientific names from Munz and Keck (1968); species codes in short names from Powell (1987): see Appendix 1. Cover types can be grouped in six series on the basis of the dominant oak species, and a seventh series for mixed dominance (3 or more codominant oaks sp.).

The VTM survey plots were 0.081 hectare rectangles chosen to be representative of the vegetation map units being delineated in the field on topographic maps (Wieslander, 1935). Vegetation information consisted of the number of stems per diameter class of overstory tree species and the percent cover of each understory plant species. Original environmental information included plot location, elevation, slope, aspect, soil surface texture and parent material.

Although the survey does not cover the entire distribution of oak species in California (in particular in the north and northwest), it provides an adequate coverage of the current range of the main oak species (Fig. 1). Temperature, length of growing season and evapotranspiration data for each plot were derived by Evett (1994) from a series of county level isoline maps drawn by C. R. Elford, the California state climatologist, in the 1960 s and 1970 s. These maps have been published by University Extension and various state agencies and are available in the Water Resources Library at U.C. Berkeley. We derived precipitation data from an isohyetal map of thirty-year annual average precipitation (1950 to 1980) drawn from approximately 4100 stations by the California Department of Water Resources. We computed average available soil water capacity from the State Soil Geographic (STATSGO) GIS soils database for California (USDA Natural Resources Conservation Service, National Cartography and GIS Center, P. O. Box 6567, Fort Worth, TX 76115). Both precipitation and soil map data were at 1:250,000 scale. We derived climatic zonations from (1) the generalized plant climate map of California [scale ca. 1:1,440,000], compiled by Sunset Books from Elford and coworkers 'plant climate maps and published by Pacific Gas and Electric Co., San Francisco, CA. (1989); and (2) a map of modified Koeppen climates [scale ca. 1:2,000,000] drawn by James (1966). We derived ecological zonation from a map of ecological units of California [scale 1:1,000,000] by Goudey and Smith (1994). Values for individual plots were interpolated and/or extracted from the maps using the GIS program Arc/Info (ESRI, Redlands, CA.). "Great circle" distance to the closest point on the coast was computed for each plot using Arc/Info and a FORTRAN program based on equations from Robinson et al. (1978). Potential cloudless solar radiation on a tilted surface (in MJ/m^2) was computed as a function of latitude, altitude, slope and aspect using a FORTRAN program written by T. Rumsey (personal communication) based on formulas given by Duffie and Beckman (1980). The variables used in the analyses are summarized in Table 2. A GIS database of California s hardwood rangelands developed by Pillsbury et al. (1991) and revised by Pacific Meridian (1994) using remotely sensed data for the California Department of Forestry was used to validate one of our analysis.

Table 2: List of variables used in the analyses.

Variable	*Type*	*Description*
Physiognomy:		
O_TREE	R	Total overstory tree basal area (sq. ft / acre).
U_TREE	R	Total understory tree cover (%)
SHRUB	R	Total shrub cover (%)
HERB	R	Total herb cover (%)
GRASS	R	Total grass cover (%)
GROUND	R	Total ground (i.e. litter, rock and bare soil) cover (%).
Abiotic factors:		
ALTITUDE	R	Elevation (m)
PRECIP	R	Mean annual precipitation (mm)
MAT	I	Mean annual temperature (C)
JAMI	I	Mean minimum temperature in January (C)
JAMA	I	Mean maximum temperature in January (C)
JAMEAN	I	Mean temperature in January (C)
JARAN	R	January temperature range (C) i.e. JAMA – JAMI
JUMI	I	Mean minimum temperature in July (C)
JUMA	I	Mean maximum temperature in July (C)
JUMEAN	I	Mean temperature in July (C)
JURAN	R	July temperature range (C) i.e. JUMA – JUMI
TEMPR	R	Annual temperature range (C) i.e. JUMEAN – JAMEAN
JUNRAD	R	Potential (cloudless) solar radiation on the average day in June (MJ/m^2)
DECRAD	R	Potential (cloudless) solar radiation on the average day in December (MJ/m^2)
SOLRAD	R	Yearly potential (cloudless) solar radiation (MJ/m^2)
SLOPE	O	Percent slope class (0-15, 16-25, 26-35, … , > 86)
ASPECT	N	Exposure [1=N, 2=NE, 3=E, ...9=flat)
AWCL	R	Available water capacity of component with lowest value in soil map unit (mm)
AWCH	R	Available water capacity of component with highest value in soil map unit (mm)
AWCAVG	R	Average (over components) of available water capacity for soil map unit (mm)
COASTDST	R	Distance to the closest point on the coast (km)
CLIMZONE	N	Plant climate zone (19 classes).
KOEPPEN	N	Modified-Koeppen climate zone (11 classes).
ECOREG	N	Ecological unit, section (13 classes).

Type: N for nominal, O for ordinal, I for interval and R for ratio (following Stevens, 1951).

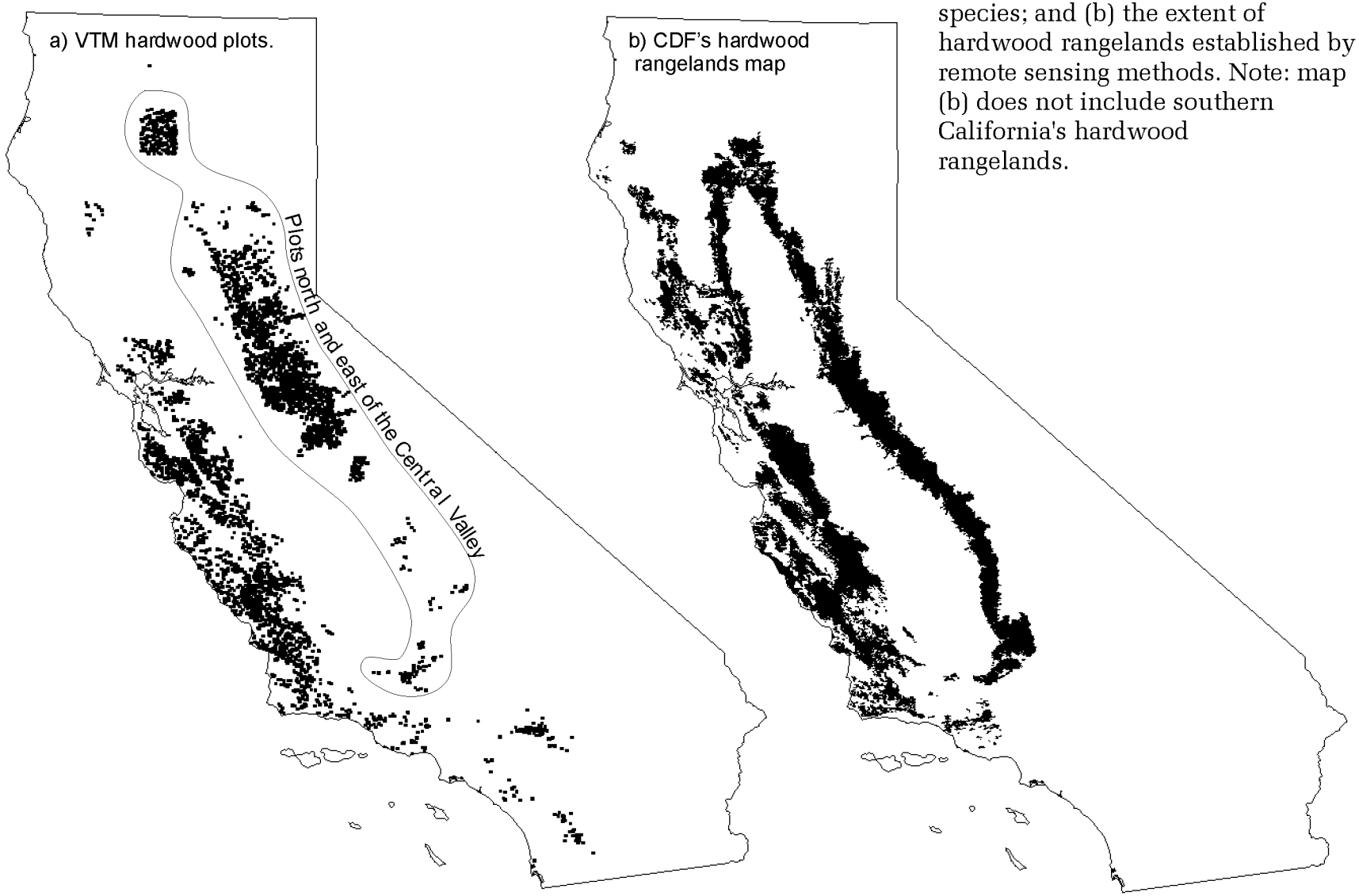


Figure 1: Maps of (a) the 4288 VTM survey plots featuring a Quercus species; and (b) the extent of hardwood rangelands established by remote sensing methods. Note: map (b) does not include southern California's hardwood rangelands.

Supervised conceptual clustering.

Clustering and classification.

Clustering and classification are basic scientific tools used to systematize knowledge and analyze the structure of phenomena. Both refer to the process of partitioning a set of objects into groups such that the members of each group are as similar as possible to one another and the different groups are as dissimilar as possible from one another. Unfortunately, although some basic distinctions in this process are recognized across disciplines, common terminology is lacking, and the two terms are often used interchangeably. Since our approach does not fit squarely in either of the conventional types of analysis, we need to clarify the context as a first step in exposing our methods.

The conventional distinction made between clustering and classification is the following. Clustering is the process of partitioning a set of items (or grouping individual items) into a set of categories. Classification is the process of assigning a new item or observation to its proper place in an established set of categories (Anderberg, 1973). In clustering, little or nothing is known about the category structure, and the objective is to discover a structure that fits the observations. In classification, the category structure is known a priori, and the objective is to recognize a new observation as a member of one of the categories. In other words, the aim of clustering is descriptive whereas that of classification is predictive. At the procedural level, clustering does not differentiate among attribute variables. It uses all attributes both to define a measure of the similarity/dissimilarity among objects and groups (i.e. an objective function), and to delineate the groups that optimize the objective function. In clustering, the goal of the grouping process is intrinsic to the set of attributes. On the other hand, classification uses one of the attributes variables (a given partition) to define the objective function, and the other attributes to delineate the groups. Thus, the goal of the grouping is extrinsic to the set of attributes. Although the conceptual distinction we have just made is widely accepted, it is not always reflected in the terms clustering and classification. Instead, following Williams (1976), some researchers use the terms *intrinsic classification* for the case where all attributes are regarded as equivalent and *extrinsic classification* when they are not. In some fields of study, the term *unsupervised classification* is used to indicate the absence of prior knowledge of groupings (i.e. for clustering), while s*upervised classification* is used to stress that

the class structure is known a priori. In machine learning, researchers make the same basic conceptual distinction under the terms unsupervised learning or learning from observation, and supervised learning or learning from example. We propose to use the qualifier *unsupervised* for problems where the goal is intrinsic, and the qualifier *supervised* for problems with extrinsic goals (columns of Table 3).

Table 3: Typology of classification and clustering problems based on goal and assessment.

	INTRINSIC GOAL All attributes are used to define the objective function and to delineate the groups.	***EXTRINSIC GOAL*** One (or more) attribute is used to define the objective function, the others to delineate the groups.
INTRINSIC ASSESSMENT The resulting groups are of interest in their own right.	UNSUPERVISED CLUSTERING a.k.a. clustering, intrinsic classification, unsupervised classification, learning from observation.	SUPERVISED CLUSTERING
EXTRINSIC ASSESSMENT The resulting groups must reflect some reference grouping	? Logically should be called unsupervised classification, but term already used.	SUPERVISED CLASSIFICATION a.k.a. classification, extrinsic classification, learning from example.

There is another, more essential distinction between clustering and classification that hinges on how the groups resulting from the analysis are assessed. In the case of clustering the objective is to discover a new set of categories, thus the new groups are of interest in their own right. These groups cannot be compared to a set of reference categories: the assessment is intrinsic. As a consequence, in clustering problems group assessment is often qualitative because the value of the new categories must be ascertained in terms of how much they ease understanding, allow for the generation of new hypotheses or facilitate management. In the case of classification the objective is to recognize a new

observation as a member of one of several given categories. Therefore, the groups resulting from the analysis must reflect some reference set of categories: the assessment is extrinsic. In classification problems group assessment can be reduced to a statistic: the prediction error rate or misclassification rate. This measures the proportion of objects (for which the reference grouping is known) that are allocated to the wrong group. It is generally assumed that intrinsic assessment is always associated with an intrinsic goal, and that an extrinsic assessment necessarily follows an extrinsic goal. We argue that this is not necessarily so (Table 3), which allows us to distinguish *supervised clustering* analyses, where the goal is extrinsic but the groups 'assessment is intrinsic. The conventional types of analysis, where goal and assessment are either both intrinsic (unsupervised clustering) or both extrinsic (supervised classification), are also found in Table 3. A fourth type, with intrinsic goal and extrinsic assessment, should be named *unsupervised classification* but this term is already commonly used for unsupervised clustering. Although this fourth type is conceivable, we have not found practical justifications for its use.

Strategies and algorithms.

In terms of strategy the choice is between *hierarchical* and *non-hierarchical* methods (Williams 1976). The non-hierarchical strategy optimizes the individual groups, which are made as homogeneous as possible. The hierarchical strategy optimizes a route between the entire population and the set of objects of which it is composed. This route may be defined by progressive fusions or by progressive divisions. *Agglomerative* methods start with single objects and recursively combine them into groups, by fusion of objects or groups into larger groups. *Divisive* methods start with all objects in the population as a group and proceed by successive divisions until all groups contain only a single object or some kind of stopping rule is satisfied. Also of interest for our discussion is the distinction between monothetic and polythetic algorithms. An algorithm is said to be *monothetic* when its clustering steps are based on one variable at a time. Algorithms that use all variables simultaneously are called *polythetic*. Polythetic methods may identify multivariate structures in *p*-dimensional space that may not be captured by a monothetic method. On the other hand, the relative simplicity of the monothetic approach make it better able to deal with large data sets and its results are much easier to interpret because we know which variables caused each clustering step (or each split, in the case of divisive algorithms).

Conceptual clustering vs. conventional clustering.

All clustering methods require a measure of group homogeneity. Most conventional cluster analysis methods are polythetic and define clusters on the basis of the pairwise similarity (or dissimilarity) of objects. They reduce the similarity between any two objects to a single number: the value of a particular function applied to the multivariate description of objects. Typically the similarity function is the reciprocal of a distance measure or metric (e.g. the Euclidean distance) computed from the objects 'vectors in attribute space. Anderberg (1973) and Mirkin (1996) provide general reviews of similarity/dissimilarity measures, and van Tongeren (1995) discusses the similarity indices often used in plant ecology.

The use of distance metrics for clustering has some drawbacks, however. First, distance values are laden with assumptions. Even when all variables are measured on a real number scale, computing a distance is not straightforward. Any transformation, standardization or weighting of the data can change the relative importance of individual attributes and thus influence the resulting classification. When data attribute types are mixed (i.e., some combination of binary, categorical and quantitative variables), creating a common distance metric necessitates some ad hoc assumptions. Secondly, since dimensionality is reduced prior to the analysis, each attribute variable has the same importance in determining the groupings. As a result, distance-based methods cannot distinguish between important variables and variables that are noisy but have no connection with the underlying cause and effect that determines the bulk of the features in the dataset (Matthews and Hearne, 1991). Insignificant or spurious features have to be filtered out of the dataset in advance of the clustering process. This means that to pick the correct distance measure and weighing scheme, and select which attributes to include in the analysis, one needs to have some a priori knowledge of the target classification and of the relationships among attributes.

Methods based on numerical similarity have a third and major disadvantage: although the goal of clustering is descriptive, they do not provide an explicit description of the clusters. The groups are defined by the list of individual objects they contain and do not necessarily have any simple conceptual interpretation. In fact, in any cluster some objects will possess and some will lack any attribute (Macnaughton-Smith, 1965). The description of the groups in terms of the original attributes (or some other criterion) is left to the interpretation of the researcher.

This also complicates the classification of new objects: the appropriate distance from the new object to each existing group must be computed before it can be allocated to one of them. Often the researcher has to use some other method to build a decision key to describe the groups and classify new objects. Such lack of "meaning" of the groups in terms of the concepts typically used by human for classification has fostered the development of conceptual clustering methods.

There are two different ways to define a group of objects: (1) by *extension*, that is, by enumeration of all the objects it contains; or (2) by *intension*, that is, by a meaningful description involving its attributes or features (Mirkin, 1996). *Conceptual clustering* refers to any method that determines a structure in a collection of objects, in which the nodes represent concepts and the links represent relationships between the concepts (Michalski and Stepp, 1982). Each cluster is defined not only extensionally but also intensionally by its corresponding concept. The concepts often take the form of a conjunctive statement involving the attributes of the set objects. For example the concept "burnt, steep-slope, chamise chaparral" uses a conjunction of classes or intervals, namely disturbance, slope, dominance, physiognomy, to define a landscape unit. The roots of conceptual clustering can be traced back to association analysis in plant communities (Williams and Lambert 1959, 1960) a monothetic, divisive method. Kaufman and Rousseeuw (1990) have implemented a variant of this method based on the chi-squared statistic. Work on monothetic divisive predictive algorithms dates back to predictive attribute analysis (Macnaughton-Smith, 1963) and was pursued by Sonquist et al. (1973).

Most conceptual clustering techniques use monothetic divisive algorithms because of the following advantages. Divisive methods are less likely to misrepresent the main structure of the data than agglomerative methods which start with single individuals and may suffer from undue influence of low level features (Williams, 1976). Monothetic methods allow for variable selection and the detection of interactions among variables (Sonquist and Morgan, 1964). And, importantly, monothetic divisive algorithms result in structures called decision trees which facilitate the intensional definition of clusters and allow for the quick and unambiguous allocation of new objects. The difficulty of these methods is that considering all possible divisions of the data into two subsets involves a very large number of combinations. Therefore, until recently, the computational requirements have been prohibitive. This is why divisive algorithms have been

largely ignored in the literature and most published methods of cluster analysis are agglomerative (Kaufman and Rousseeuw, 1990). The increasing availability of computing power has given a new impetus to monothetic divisive methods in the 1980's with contributions from statisticians searching for efficient nonparametric methods (Breiman et al., 1984) and from the artificial intelligence community seeking techniques for machine learning (Quinlan, 1986).

Supervised conceptual clustering.

The traditional distinction between intrinsic and extrinsic goals is also valid for conceptual methods. Many authors use it to distinguish between two types: conceptual clustering methods and concept classifiers (e.g. Briscoe and Caelli, 1996). However such a distinction is not as clear-cut as often thought. A more general classification of conceptual methods in three types, each tracing back to the early 1960's, shows that these methods are in fact part of a continuum. Borrowing from Mirkin's (1996) typology, we sort methods according to "classification learning task":

(a) Self-learning: the method must provide a conceptual structure for a given set of variables, without a priori knowledge of structure. Variables are used both to define the goal of the grouping process (i.e. a utility function to optimize), and to delineate clusters and define concepts. This task corresponds to the first column of Table 3. Methods include association analysis (Williams and Lambert, 1959), CLUSTER/2 (Michalski and Stepp, 1983), COBWEB (Fisher, 1987) and RIFFLE (Matthews and Hearne, 1991; Matthews et al., 1995).

(b) Learning a partition: the method must provide a decision rule for distinguishing between the classes of a given partition. One variable is used to define the goal of the grouping process (here, a goodness-of-split criterion to be maximized), while the others are used to delineate clusters and define concepts. This task corresponds to the second column of Table 3. Methods include predictive attribute analysis (Macnaughton-Smith, 1963), AID3 (Sonquist et al. 1973), THAID (Morgan and Messenger, 1973), CHAID (Kass 1980), CART (Breiman et al. 1984), ID3 (Quinlan, 1986) and C4.5 (Quinlan, 1993).

(c) Learning the association between two sets, $\boldsymbol{X}$ and $\boldsymbol{Y}$, of attribute variables: the decision tree is designed using $\boldsymbol{X}$-variables to produce clusters that are meaningful with regard to $\boldsymbol{Y}$-variables. The $\boldsymbol{Y}$-variables are used to define the utility function to be optimized and $\boldsymbol{X}$-variables are used to delineate clusters

and define concepts. An early example of a method designed to accomplish such a task is multiple predictive analysis (Macnaughton-Smith, 1965). Another method, due to Rostovtsev and Mirkin (1985) is cited and succinctly presented in Mirkin (1996). These two methods differ in the way they process the ***Y***-variables: multiple predictive analysis combines the ***Y***-variables into a single outcome vector; Rostovtsev and Mirkin sum up the measures of dispersion computed for each of the ***Y***-variables into a single goodness-of-split criterion.

As noted by Mirkin, tasks (a) and (b) are both particular cases of a more general task (c): self-learning is a case where ***Y*** = ***X***, and partition learning a case where ***Y*** is reduced to a single categorical variable representing the given partition. Fundamentally, all grouping problems are about discovering associations among the variables at hand. Whereas the self-learning (task a) methods fit the clustering designation and partition-learning (task b) methods fit the classifiers designation, association-learning (task c) methods does not fit either. In fact, task (c) methods may gravitate to either category depending on the respective number of variables in the ***X*** and ***Y*** sets.

Mirkin also points out that, in many cases, methods designed to learn a given partition (task b) are used to find a "theoretical", intensional cluster structure to describe an "empirical", extensional one that is given. In such a case, a "classification" method is in fact used for "clustering". We argue that whether the task (b) method is used for classification or clustering hinges on whether the new groups 'assessment is intrinsic or extrinsic. In other words, it hinges on the way one treats the discrepancy between the given (extensional) structure and the learned (intensional) structure. Discrepancy may spring from various sources: from an incomplete or inadequate set of attributes; from limits of the method (none of the decision tree methods guarantees to find the absolute best partition); and, in real-world datasets, from the presence of some amount of noise. Within a classification approach, objects belonging to one of the given classes that are allocated to another class are considered as errors in prediction: they are misclassified. The newly derived intensional structure is used to describe the given partition and to allocate new cases to its classes. The quality of the intensional structure can be assessed by its prediction error rate. Within a clustering approach, objects belonging to one of the given classes that are allocated to another class are not considered as errors: they are reclassified. The newly

derived intensional structure is related to the given partition, but it provides a different information and will serve different purposes. Thus the prediction error rate cannot be used to assess the quality of the intensional structure, although (1 - prediction error rate) may be useful as an indicator of the strength of the relationship between the learned partition and the given partition. The idea of using a partition-learning (task b) decision tree method for clustering can be traced back to Anderberg (1973) who envisaged using AID for "clustering with respect to an external criterion". We prefer the more concise "supervised clustering" to stress that the clustering process is directed by some a priori knowledge of structure. To our knowledge there is no elaboration or documented use of such an approach in the literature.

Before describing our supervised clustering analyses in more details, we need to present the partition-learning method we have used and discuss some of its characteristics.

CART: our choice of recursive partitioning method.

We chose to use classification and regression trees (Breiman et al., 1984). CART belongs to the family of divisive monothetic methods, generating decision trees from a set of learning cases. It operates by recursive partitioning, splitting the initial set into subsets that are more homogeneous in terms of the outcome variable (i.e. the given partition). Each of the successive splits is made at a particular value of an attribute variable. Decision trees are branching diagrams that are similar in concept to dichotomous taxonomic keys. They are typically used to allocate new objects to the classes of a given partition: starting from the root node, a question is asked at each node of the tree and data cases for which the answer is yes are assigned to one branch while the others go to the other branch. Logical combinations of the answers to these questions can also be used to define a conceptual description for each of the terminal nodes. The terminal nodes (leaves) are the clusters constituting the end product of the partition learning task. The reasons that lead us to choose CART after trying another decision tree program are the following. CART offers effective goodness-of-split criteria, it includes a very efficient way to deal with overfitting, and it can use any combination of categorical and continuous variables. Also, usable PC-based implementations were available at the time of our study: CART for Windows version 2.0 and CART for DOS version 1.01 developed by Salford systems Inc. (Steinberg and Colla, 1995).

The two fundamental problems in constructing an effective decision tree are finding good splits and limiting the size of the tree to avoid over-fitting the data. At each step, the CART procedure (1) considers all possible splits for all variables; (2) ranks each splitting rule on the basis of a goodness-of-split criterion reflecting the degree of homogeneity achieved in the child nodes; (3) chooses the splitting rule that maximizes the goodness-of-split criterion to separate the node's objects in two subsets. This process is repeated recursively until further splitting becomes impossible: i.e. when only one case (or a preset number of cases) remains or when all the node's cases belong to the same class. Typically, end-nodes (the leaves of the tree) are classified as the plurality of the learning cases they contain. However, other classification rules can be used. Rather than using a stopping rule to decide when to stop growing the tree, CART grows the largest possible tree and then enters a process of elimination of superfluous branches that the developers of CART call "pruning back to an honest tree" (Breiman et al., 1984).

Goodness-of-split criteria.

Different splitting criteria can be used to grow decision trees. Many programs (e.g. ID3) use the entropy criterion: an impurity measure based on information theory. For multiclass problems, the CART program offers a choice between the gini and twoing criteria (Breiman et al., 1984; Breiman 1996). These criteria are best presented in mathematical form. Let the target partition be given by the variable Y taking on values $j = 1, \ldots, J$. Assuming that a set S of splits at every node t has been specified, splitting rules are defined by specifying a goodness-of-split function $\theta(s, t)$, defined for every $s \in S$ and node t. At every t, the split adopted is the split s^* which maximizes $\theta(s, t)$. If $\boldsymbol{p} = (p_1, \ldots, p_j)$ are the proportions of the J classes in t, then $\phi(\boldsymbol{p})$ is an impurity function if it is convex in $\boldsymbol{p}$, has a maximum when all p_j are equal and is minimum when one of the $p_j = 1$. A split s divides t in two child nodes t_L and t_R sending a proportion P_L of the objects in t to t_L and a proportion $P_R = 1 - P_L$ to t_R . The proportions of the J classes in the child nodes are $\boldsymbol{p}_L = (p_{1,L}, \ldots, p_{j,L})$ and $\boldsymbol{p}_R$ respectively. For the impurity function $\phi(\boldsymbol{p})$ the goodness-of-split is defined as:

$$\theta(s,t) = \phi(\boldsymbol{p}) - P_L\phi(\boldsymbol{p}_L) - P_R\phi(\boldsymbol{p}_R)$$

The gini diversity index is an impurity function that has the form:

$$\phi(\boldsymbol{p}) = \sum_j p_j(1 - p_j)$$

for example, for the node distribution (1/4,1/4,1/2), the gini diversity index is:

$$1- (1/16 + 1/16 + 1/4) = 1- 0.374 = 0.626$$

The gini index can be interpreted as a measure of qualitative variance. Its value is maximum for a node with equal proportions of each class (e.g. $\boldsymbol{p}$ = (1/3, 1/3, 1/3)), and zero for a node containing only one of the classes (e.g. $\boldsymbol{p}$ = (0,0,1)). The best gini splits try to produce pure nodes, i.e. sending all data in the class with the largest p_j to t_L and all other classes to t_R.

The twoing criterion finds the particular grouping of all J classes in two superclasses that results in the greatest decrease in node impurity when the two superclasses are treated as a two-class problem. If the gini impurity measure is used in the two-class problem, then the best twoing split at a node maximizes

$$\theta(s,t) = \frac{P_L P_R}{4}\left[\sum_j \left| p_{j,L} - p_{j,R} \right|\right]^2$$

and the two superclasses corresponding to the split maximizing θ are

$$C_1 = \left\{ j; p_{j,L} \geq p_{j,R} \right\} \text{and } C_2 = \left\{ j; p_{j,L} < p_{j,R} \right\}$$

The twoing approach was designed to give strategic splits and inform the user of class similarities. At each nodes it sorts the classes into the two groups that are, in some sense, most dissimilar. Near the top of the tree, the twoing criterion attempts to group together large number of classes that are similar in some characteristic. Near the bottom of the tree it attempts to isolate single classes.

In their 1984 book, CART developers had reached the general conclusion that the properties of the final tree were surprisingly insensitive to the choice of splitting rule, and that the criterion used to prune the tree (or recombine nodes upward) was much more determinant. Later, Buntine and Niblett (1992) compared the gini index, the entropy or information gain (Quinlan, 1986), the Marshall correction (Mingers, 1989) and a random split selection, using Breiman et al. ś pruning method with each of them to remove the effect of the pruning criterion on the final tree. They found the gini and information gain criteria to be the best in terms of classification accuracy.

Breiman (1996) revealed interesting differences between the best splits selected with the gini and twoing criteria. The gini criterion favors splits that put the largest class into one pure node, and all the others into the other node. Twoing (as well as entropy) favors splits that tend to balance the sizes of the two children nodes, i.e. the split that minimizes $|P_L - 0.5|$. When Y has a small number of classes, one can expect both criteria to produce similar results. It is with larger J

that differences will become apparent. When dealing with many classes, gini may produce splits that are too unbalanced, especially at the root of the tree. However, twoing and entropy suffer from a more disturbing problem. As J grows from moderate to large (say $J \geq 10$), there are usually many combinations of the J classes in two superclasses such that $P_L \approx 0.5$. Thus, selecting the best split with these criteria becomes a bit arbitrary. Breiman, suggested that, in such conditions, twoing and entropy should be combined with a limited two-step look ahead algorithm.

Pruning trees to avoid over-fitting the data.

Many clustering and classification methods are assuming self-contained (i.e. completely enumerated) populations, and error-free attribute information. However, real-world data sets are often population samples that invariably contain noise, that is, non-systematic errors in attribute values and/or object misclassifications. Given adequate attributes, basic decision tree algorithms typically split the dataset until perfect homogeneity is achieved, even if this means that each leaf node contains only a single object. But these large trees do not generalize well to new samples from the same population because they include noisy structures that are specific to the learning sample. This tendency of models to overfit sample data is well known to statisticians.

To control overfitting, CART developers introduced tree pruning: a major improvement on earlier methods (e.g. AID, CHAID, ID3) which used various stopping criteria to decide when to stop growing a tree. First CART grows the "maximal tree", then it evaluates smaller trees obtained by pruning away branches, and retains the best subtree. During this process, each subtree is tested for its error rate or "misclassification cost" on data that were not used to grow the maximal tree. Two approaches can be used: (1) if the data are many, the set is divided into separate learning and test subsamples; (2) if data are few, CART uses cross-validation (usually ten-fold). Cross-validation mimics the use of a test sample while extracting information from all the cases of a dataset to develop the model. Tenfold cross-validation divides the data in ten subsamples of equal size and with similar proportions of the response variable outcomes. Each of the ten subsamples is set apart in turn and serves as a test sample for the other nine. During each run, ninety percent of the data serves to build trees that are tested on the remaining ten percent. The total misclassification cost, computed over the ten runs, serves to

determine the best tree size. The final tree is constructed from all of the data, using the best tree size.

Choice of priors.

In many studies, the various classes composing the target partition are not equally represented in the sample data. Since CART minimizes the overall misclassification rate, the best tree may predict numerous classes with good accuracy while grossly misclassifying classes with small number of cases. In other words, it is less costly to misclassify a member of a small class than to misclassify a member of a numerous class. Priors can be specified as a parameter to adjust class misclassification rates in any desired direction. For instance, equal priors will tend to equalize class misclassification rates. Prior probabilities represent the probability of observing a particular class in the population and they are to override the proportions present in the sample. In CART, the goodness-of-fit function and within-node probabilities are adjusted using the prior probabilities specified at the beginning of the analysis. In that case, end-node class assignments do not follow simple plurality, but prior-weighted plurality, and the misclassification rates are also weighted.

Supervised clustering analyses.

Now that we have discussed the methodological context and the method used for our analyses, we present in more details what we have called supervised clustering. To be more concrete, we will use specific analyses in this study to illustrate 3 cases along a spectrum from supervised classification to supervised clustering: (1) learning an intensional structure for a given classification; (2) simplification-generalization of a given classification; and (3) supervised clustering. All three analyses are cases of Mirkin ŝ learning task (c): learning the association between two sets, ***X*** and ***Y***, of attribute variables.

The first analysis is reported mainly to clarify the context of the other two and to illustrate one end of the spectrum. In this case we used individual species abundance data to predict the hardwood cover type of the 2038 VTM plots classified by Allen et al. (1991). The ***Y*** variable (the predicted variable for CART) is the cover type, with 57 possible values (Table 1). The group of ***X***-variables (the predictors) comprises the 170 plant species most often found in the survey plots, with values of cover or basal area. Although the cover types contain some

environmental information, they were derived mostly from the relative abundance of the species, using TWINSPAN. Therefore, in this case, it is fair to say that $\boldsymbol{Y} = \boldsymbol{X}$, or at least that the $\boldsymbol{Y}$-variable and $\boldsymbol{X}$-variables contain the same information. The benefits of such an analysis are the following. (1) Because CART goes to a process of variable selection, the resulting tree provides a decision key to classify new plots with a much-reduced set of species. Its prediction error rate indicates how well it can classify new plots among the given classes. (2) The tree also provides a basis for conceptual descriptions of the types (although TWINSPAN indicator species already provided that in this particular case). The prediction error rate measures how well the new structure fits the given one. Because predicted variable and predictor variables are two expression of the same information, this first analysis is a particular case of supervised classification. It amounts to finding an intensional cluster structure to describe an extensional one that was derived by other methods. It is a true classification procedure, but used in a clustering perspective.

In the second analysis we use CART to predict the cover types of 1992 VTM plots (as classified by Allen et al.) from life-form abundance data. The $\boldsymbol{Y}$ variable is the plot cover type, with 57 possible values (Table 1). The $\boldsymbol{X}$-variables group comprises six variables (Table 2), summarizing abundance (cover or basal area) values by plant life forms (overstory tree, understory tree, shrub, herb, grass), and ground (litter, rock, bare) cover. While the cover types were derived from relative species abundance, they also reflect plot physiognomy (i.e. the relative abundance of the different life forms). This is apparent in the cover type denominations (Table 1), e.g. "*Quercus agrifolia* / Grass" reflects a different plot physiognomy than "*Quercus agrifolia / Holodiscus discolor-Symphoricarpos rivularis*." Plot physiognomy was included in part through the use of pseudo species in the TWINSPAN analyses, with cut values of 2, 5, 15, 25, 75, and 95. The analysts also introduced physiognomy by not treating all life forms the same way: overstory trees as a group were given particular importance since they are the "dominant" species, all grass species were combined into a single "species", etc. Therefore, in this case it is fair to say that $\boldsymbol{X} \subset \boldsymbol{Y}$, that is, the $\boldsymbol{X}$-variables contain a subset of the information contained in the $\boldsymbol{Y}$-variable. The decision tree provides a reclassification of the original plots into new clusters. The new partition should be simpler than the original one (the cover types) because information about individual species is not included in the $\boldsymbol{X}$-variables. However, because CART is

trying to predict the cover type, the new partition retains some species-level information. There are many possible ways to distinguish clusters of plots using life form abundance (the ***X***-variables) only. A self-learning method (task a) would have done so by optimizing a function of these ***X***-variables. By trying to predict cover types (the ***Y***-variable) from life form abundance, we have chosen a particular set of clusters: those that retain the most information about cover types and thus about relative species abundance. In this case, the prediction error rate given by CART is not interpreted as a measure of misclassification, but rather as an indication of how much of the original information is retained in the new classification.

In the third analysis we use CART to predict the physiognomic groups (identified in the second analysis) from abiotic factors reflecting geographical position, topography, soil and climate. The ***Y*** variable is the plot's physiognomic group, with 26 possible values (Table 5). The ***X***-variables group comprises the abiotic variables in the second part of Table 2. This case differs from the previous one in that $\boldsymbol{X} \neq \boldsymbol{Y}$, that is, the ***X***-variables do not carry information already included in the ***Y***-variable. Indeed, most of the abiotic data we use were not available at the time of the classification in cover types. This case is clearly a relationship learning task (task c). If there is no relationship between ***Y*** and ***X***, CART will not grow a tree. By trying to predict physiognomic groups (the ***Y***-variable) from abiotic variables, we are again choosing a particular set of plot clusters: those that maximize the relationship between physiognomic groups and abiotic factors. Here again, the prediction error rate given by CART is not interpreted as a measure of misclassification, but rather as an indication of the strength of the relationship between ***Y*** and ***X***. However, this extended use for the misclassification rate is not to be pursued beyond the choice of a particular CART model. A true assessment of the new clusters will require other means.

Assessing the new classifications.

We have seen that the clustering approach is characterized by intrinsic assessment, that is, the resulting groups are of interest in their own right, not as an attempt to reproduce a reference grouping of the data. Consequently, whereas a single statistic can express the goodness-of-fit to a reference grouping, without such a reference there is no criterion for judging the "goodness" of clusters. There might be a strong temptation to apply ANOVA or standard tests of significance on

group differences to test for significance of the structure. The trouble with making such tests is that they are hardly relevant (Anderberg, 1973). Since the clustering method was designed to produce clusters that are well differentiated from each other, "highly significant" results should occur with monotonous regularity if testing is done on the variables used in the clustering (Milligan, 1996). Finding a useful criterion to measure cluster "goodness" is very unlikely for similar reasons. Any criterion designed to measure cluster separation (or cluster homogeneity) will validate clusters built with a method having similar underlying assumptions and refute the results of methods that do not. Having a good understanding of the properties of clustering algorithms is important because to a certain extent they impose their own structure to the data (Kaufman and Rousseeuw, 1990). The difficulty is that, in most non-trivial clustering problems, there will be several different yet meaningful ways of organizing data into groups. Each of several alternative groupings may be taken as the realization of some classification principle embodied in the data and represent one of several facets of the problem (Anderberg, 1973). To assess the value of a cluster analysis, one can not depend on the mechanical workings of the algorithm but must resort to contextual knowledge, investigative purpose and interpretative skills.

If there is no "right answer" to a clustering problem, assessment of a particular grouping requires substantive interpretation. A statistician or clustering expert cannot directly perform this task; the interpretation must be based within the context of the applied discipline area of the research (Milligan, 1996). By design, conceptual clustering methods provide much help to the interpretation of the partition and its clusters. In particular, CART trees provide an intuitive representation of cluster relationships in a hierarchical structure and, at the same time, an aid to intensive description of the groups in terms of the clustering variables. However, help to interpretation should also include descriptive statistics (mean, range, quantiles, cross-tabulations) for the clusters. These descriptive measures can be computed on the variables used in the cluster analysis (the ***X***-variables) as well as for exogenous variables not directly involved in establishing the clusters. In the particular case of supervised clustering, cross-tabulating the new groups with the classes used to direct the clustering (the Y-variable) is also useful. It will provide insight about the relationship between the new groups and the given partition that goes much beyond a simple misclassification rate. Visual aids to interpretation are best since, through evolution, humans have developed

powerful classification and pattern recognition skills for arrangement in up to three dimensions. Useful graphs for summary statistics include histograms, 3D scatter plots and box plots. Frequency matrices resulting from cross-tabulations (also called contingency tables) can be described graphically in a two-dimensional graph with density plots or using correspondence analysis (Greenacre, 1993).

The use of a decision tree method such as CART for clustering adds a particular step in the interpretation of clusters. Cross-tabulating the newly obtained groups with the given classes (those used to direct the clustering) is necessary to decide whether some of the decision tree leaves need to be aggregated in a single cluster. Like most decision tree methods, CART may produce leaf nodes that are replicated, that is, leaf nodes allocated to the same original class. This occurs when two or more decision paths in the tree lead to members of a single original class. In conceptual clustering terminology, this is the way decision trees represent *disjunctive* concepts: concepts formed of two (or more) sets of conjunctive statements involving the attributes of the set objects. A landscape unit defined as "low altitude, northern exposure **or** mid altitude, southern exposure woodland" is an example of disjunctive concept. When CART is used for supervised classification, it is straightforward to aggregate the leaves predicting the same given class. But, since the leaf nodes are classified as the weighted plurality of the cases they contain, the same classification does not necessarily mean they have similar content with respect to the other classes. Also, two leaf nodes classified differently may contain proportions of the two given classes that are similar enough to justify aggregation. Thus, when using a decision tree method for supervised clustering, it is necessary to decide if the proportions of the different given classes present in the two leaf nodes are sufficiently similar to warrant aggregation. A profile (i.e. a set of frequencies divided by their total) for each cluster can be derived from the cross-tabulation of the new groups by the given classes. Such profiles are in fact the $\boldsymbol{p} = (p_1, \ldots, p_j)$ vectors used in computing the impurity functions involved in the goodness-of-split criterion.

In any data analysis, the main concern is to decide whether the results are a valid summary of the data or whether a spurious or inappropriate structure is imposed on the data. The best way to validate the analysis is to test the discovered structure on new data. We tested our physiognomic groups on 1022 VTM plots where *Quercus* species were dominant. These plots of unknown hardwood cover type were classified into physiognomic groups using the decision tree obtained

from our second analysis. We also tested some aspects of our grouping of physiognomic groups into abiotic domains by relating domains to a vegetation map established from remotely sensed data.

A final way to test the validity of a new classification is to evaluate its utility for the task it was intended for. We presented the results of the analyses to a group of hardwood rangeland experts including an academic researcher, an extension specialist, an extension agent and a private consultant. A full day workshop based on methodology developed in the field of artificial intelligence for knowledge acquisition was used to produce a S&T model for a domain. The model was transcribed as a flowchart and a catalog of states and transition, and then reviewed during a subsequent meeting. In a first step, the group used cards summarizing the physiognomy, species composition, landscape position and climatic characteristics of each of the main physiognomic groups in an abiotic domain to identify vegetation states for the domain. Then they used their experience and understanding of the dynamics of the vegetation to relate the groups with transitions.

Results.

Building a key to the hardwood cover types.

This first analysis is a case of supervised classification, we used individual species abundance data to predict the hardwood cover type of the 2038 VTM plots previously classified by Allen et al. In this case the assessment of the new grouping is extrinsic: the misclassification rate is used to measure how well the new structure fits the existing classification. As expected, using prior probabilities equal to the proportion of each cover type in the data resulted in a tree with lower misclassification rate (23% of plots misclassified when pruned to 160 leaf nodes) than using equal priors (33 % of plots misclassified when pruned to 90 leaf nodes). The corresponding un-pruned trees with 296 and 311 leaf nodes, achieved lower error rates (16 and 14%) on the 2038 data cases but did no better then the pruned trees on cross-validated data. The better overall misclassification rates reflected a much more uneven prediction of each of the cover types however. The tree based on unequal priors (set as in the data) misclassified cover types represented by many plots at a much lower rate than types with few plots. The tree based on

equal priors had misclassification rates independent of the number of plots (see examples in Table 4). Thus, it is a better predictor of the variety of cover types.

Table 4: Examples of misclassification rates depending on choice of priors.

Cover type	*n*	*Misclassification rate*	
		Priors as in data	**Equal priors**
31	291	8%	28%
17	113	6%	28%
6	23	74%	39%
52	9	100%	33%
all	2038	23%	33%

The decision tree resulting from this analysis (Appendix 2) provides a key to the cover types that is quite similar to the key constructed by Allen et al. (1991), with the first splits made on the main *Quercus* species and subsequent splits made on associated shrub species and relative levels of abundance. It was not possible to compare the two keys in terms of predictive ability because Allen et al. ŝ key is more qualitative, including qualifiers such as: usually, often, may be, sometimes and on average.

This type of analysis could also have helped define the cover types. Some paths in the decision tree constitute concepts that are almost identical to the current denomination of the corresponding cover types. For example the conjunctive concept: *Q. dumosa* basal area > 1 m^2/ha and grass cover > 27.5 % and *Q. Douglasii* basal area > 1 m^2/ha corresponds to cover type 1, named *"Q. dumosa-Q. Douglasii* / Grass*"* by Allen et al. Some paths constitute concepts that are more succinct than the cover type denomination: *Q. dumosa* basal area ≤ 1 m^2/ha and Q. agrifolia present and *Acer macrophyllum* basal area > 3.3 corresponds to cover type 51, *"Quercus agrifolia-Acer macrophyllum / Rhamnus californica-Holodiscus discolor"*. Other paths constitute concepts that require more interpretation from the analyst. For instance *Q. dumosa* basal area ≤ 1 m^2/ha and Q. agrifolia absent and *Arctostaphylos patula* cover > 8.5% correspond to type 54, i.e. *Quercus Kelloggii / Arctostaphylos patula*. Also a simple type such as type 18 (*Q. agrifolia*) may correspond to a long path in which all species often associated with *Q. agrifolia* in other types are successively eliminated.

Grouping of cover types into physiognomic groups.

This second analysis is a case of supervised clustering. We used CART to predict the hardwood cover type of 1992 VTM plots previously classified by Allen et al. from life form abundance data. However, in this case the assessment of the new grouping is intrinsic: the new structure has value in itself and the misclassification rate cannot be used to assess the clusters. The decision tree we have retained is presented in Fig. 2. The first split on grass cover (at 23.5%) separates plots where grasses are an important component of the understory to those where they are not. A second split on grass cover (at 85.5%) further separate plots where grasses almost exclusively populate the understory from those where other life forms have significant cover. Splits on overstory tree basal area separate savanna from woodland (at 14.6 –14.8 m^2/ha) and woodland from forest (at 27.4 – 31 m^2/ha). Splits on shrub cover separate plots with an understory almost exclusively occupied by shrubs (> 87.5%), dominated by shrubs (> 49.5%) or where shrubs share the understory with other life forms (> 25.5%). Splits on understory tree cover mostly separate plots that have some understory trees from those without any. The leaf nodes in the tree have been named by reference to the cover type CART predicted for that node. For instance leaf node [32] at the top of figure 2 refers to the 160 plots predicted as having cover type 32. To be consistent we will also refer to the cluster and physiognomic group corresponding to that node as [32].

The centroids of the 26 clusters corresponding to the leaf nodes of the decision tree in Fig. 2 can be found in Fig. 3. In this graph the clusters are differentiated by the three main variables (grass cover, overstory tree basal area and shrub cover) involved in the decision tree. Note that the clusters themselves appear to segregate into groups in this 3D space, which reflects the hierarchical nature of decision tree methods. Clusters whose centroids are close within these three dimensions are also in the same part of the tree and differ along another dimension. For example, clusters [2] and [52] differ in the amount of understory tree cover and the amount of ground (litter, rock, bare) cover (Fig. 4). Box plots as in Fig. 4 give a good visual representation of the differentiating characteristics (i.e. the physiognomy) of the physiognomic groups. Both centroid plots and box plots were useful to interpret the clusters as they are in the case of conventional clustering analyses. In this case they supplemented the conceptual information provided by the decision tree to help derive intensional definitions for the physiognomic groups (Table 5).

Since our clustering analysis was supervised, there is also interesting information to be derived from an examination of the relationship between the new clusters (the physiognomic groups) and the classification used to direct the clustering (the hardwood cover types). For each of the leaf nodes in the tree of Fig. 2, a profile (i.e. a set of frequencies divided by their total) can be computed from the frequencies of the different cover types present in the node. These profiles can be summarized graphically with density plots such as Fig. 5. The profile of leaf node [32], (i.e. physiognomic group [32]: "Savanna with full grass understory and some understory oaks") is the second row of Fig. 5. The largest contributor to this group is type 32 (*Quercus Douglasii* / understory *Q. Douglasii* / Grass) with 32% of the 160 plots. Next is cover type 31 (*Q. Douglasii* / Grass) with 20%, then type 8 (*Q. Douglasii-Q. Wislizenii* / Grass) and type 5 (*Q. Douglasii-Pinus Sabiniana* / Grass) with 9% each. Cover type 32 with 83% of its plots going to this node can be considered as the archetype of physiognomic group [32]. Some physiognomic groups have very similar profiles. For instance, physiognomic group [28] ("Savanna with full grass understory") has a profile that is quite similar to that of group [32]: many of its plots come from the same cover types (Fig 5) with the exception of cover type 32 who contributed none. This is consistent with the main difference between these two groups established in the decision tree: group [32] plots have some understory oaks while group [28] plots have none. The archetypal cover type for group [28] is cover type 28 (*Quercus Douglasii-Quercus lobata* / Grass) because, although not the largest contributor of plots, it has the largest proportion of its plots (75%) going to this group. Physiognomic group [24] ("Woodland with full grass understory") has also a profile that is similar to those of [32] and [28]: in this case the difference is that group [24] gathered plots with a higher overstory tree basal area. Its archetypal cover type is type 24 (*Quercus lobata-Quercus agrifolia* / Grass).

The examination of the profiles in Fig. 5 also reveals differences among physiognomic groups. For instance, group [43] ("Savanna, shrub dominated understory, some understory oaks") and group [32] gathered plots from almost completely different ranges of cover types. A synthetic way to visualize the similarity/dissimilarity of physiognomic groups in terms of cover type profiles is found in the correspondence analysis plot of Fig. 6. Axis 1 shows an opposition between physiognomic groups with an understory mostly composed of grasses and

those whose understory is mostly shrubs or litter. Axis 2 organizes the groups from the lowest overstory tree basal area (shrublands) to those with the highest (forest).

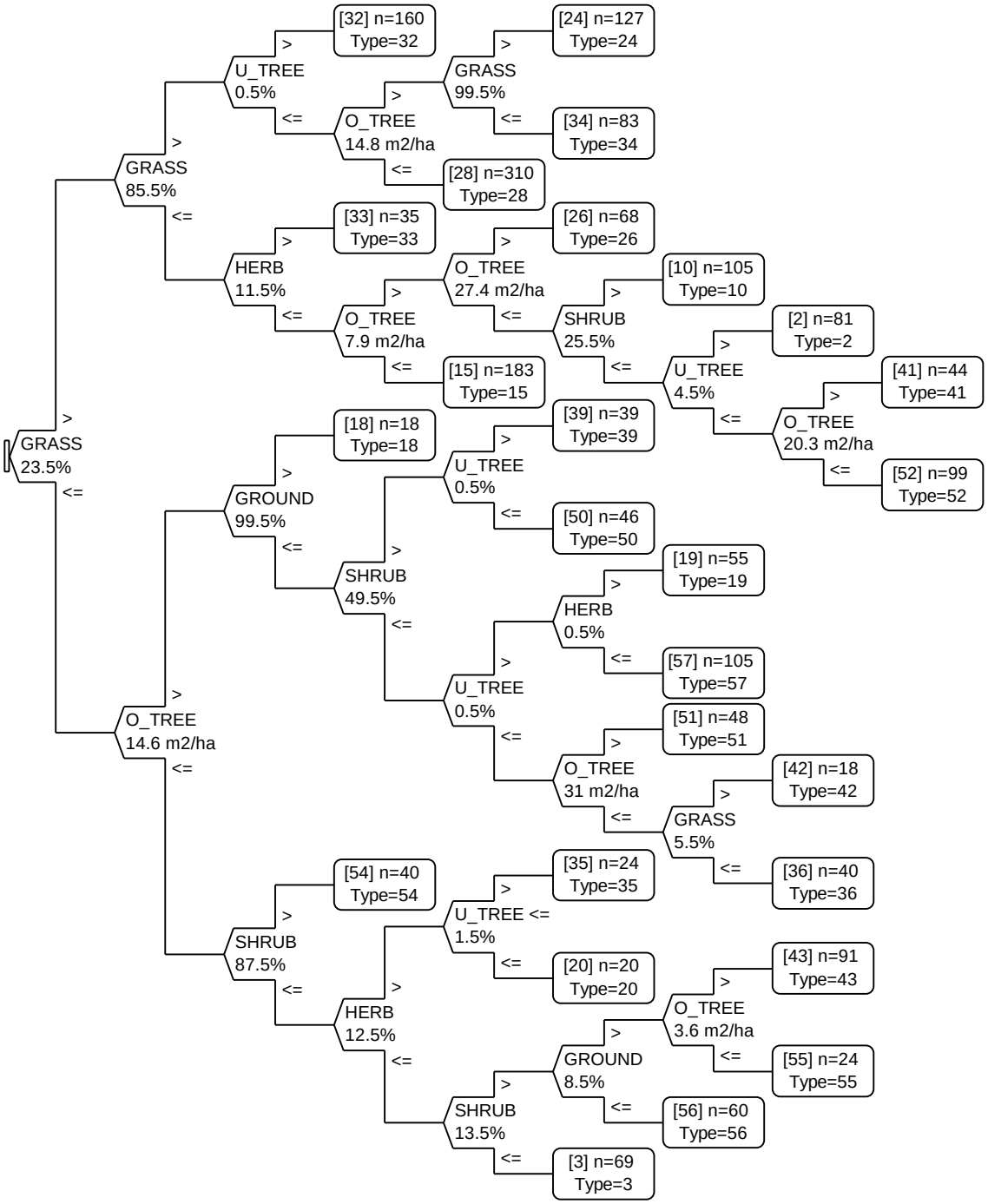


Figure 2: Decision tree of the clustering of plots into physiognomic groups. The root node of the tree is on the left and the branching nodes are marked with the corresponding split variable and split value. The leaf nodes are marked with (1) the cluster identifier, (2) number of plots in the cluster, (3) the cover type predicted for the node.

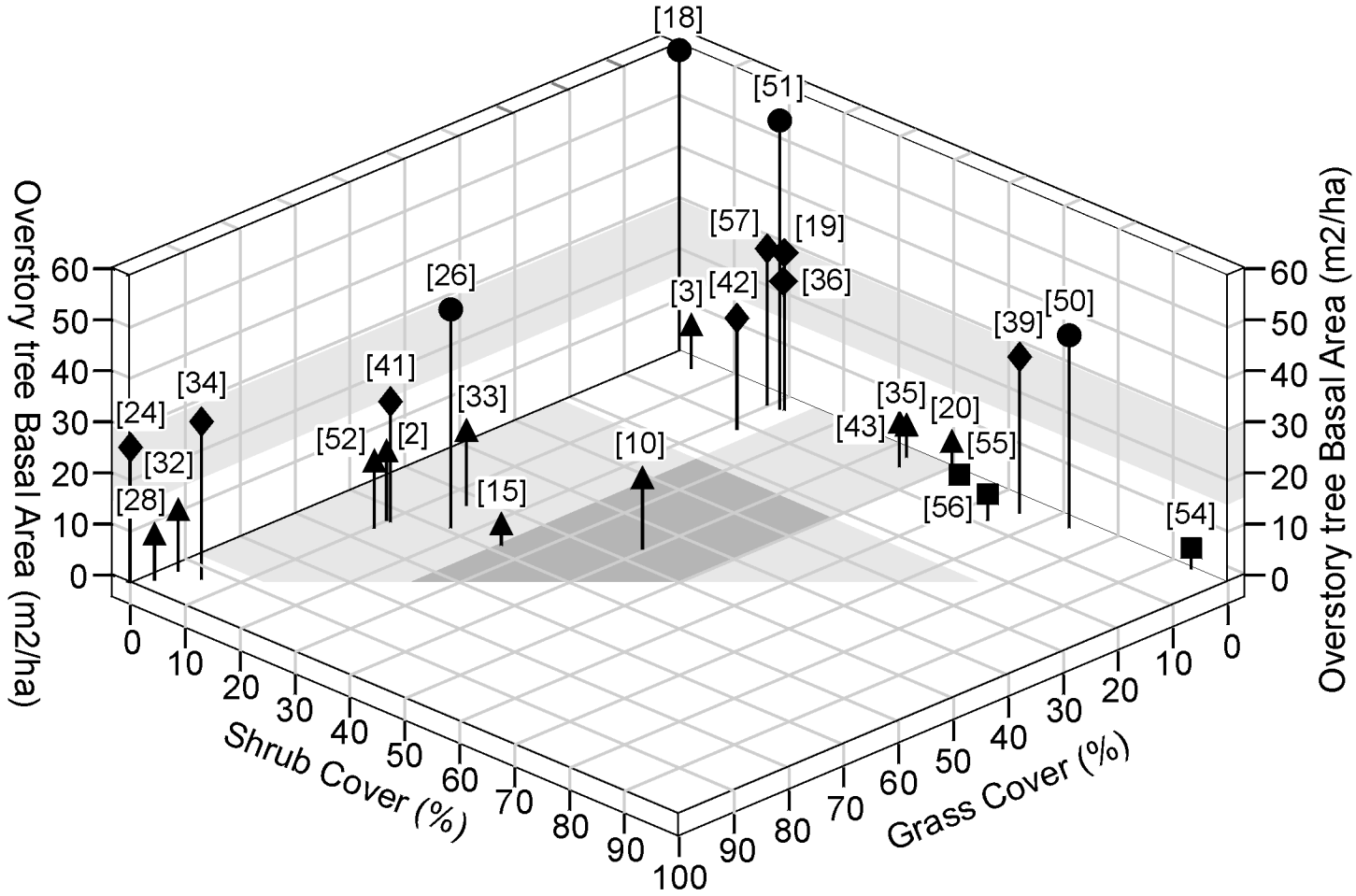


Figure 3: Cluster centroids of physiognomic groups. The shading shows variable ranges determined by the main split in the corresponding decision tree. Markers correspond to the main vegetation types of Table 5: Triangle = savanna, diamond = woodland, circle = forest, square = shrubland. Cluster identifiers are shown over the markers.

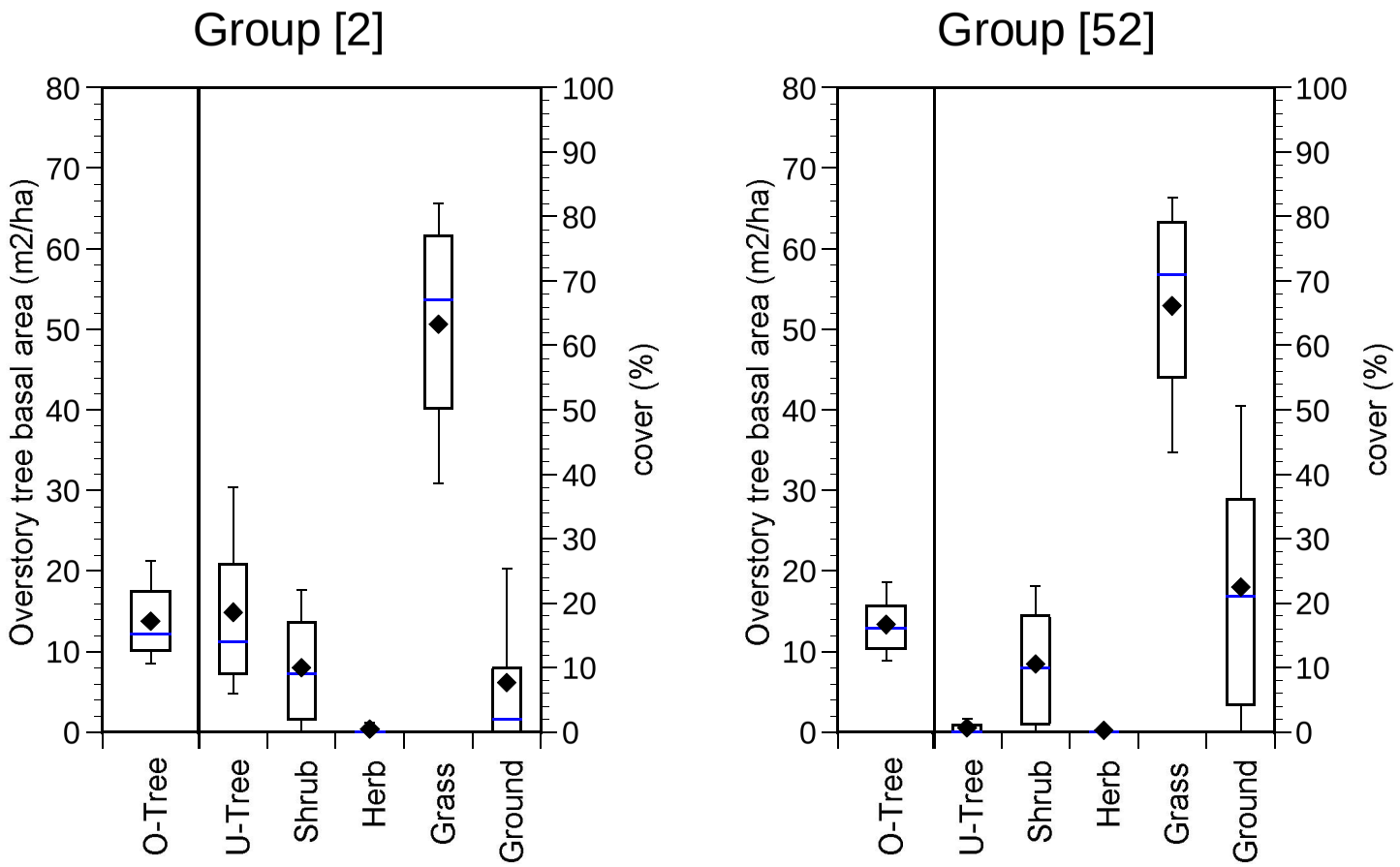


Figure 4: Physiognomic groups' differentiating characteristics. Diamond markers show the average value of plots in the group. Boxes show the range of values (10,25,50,75 and 90 percentiles).

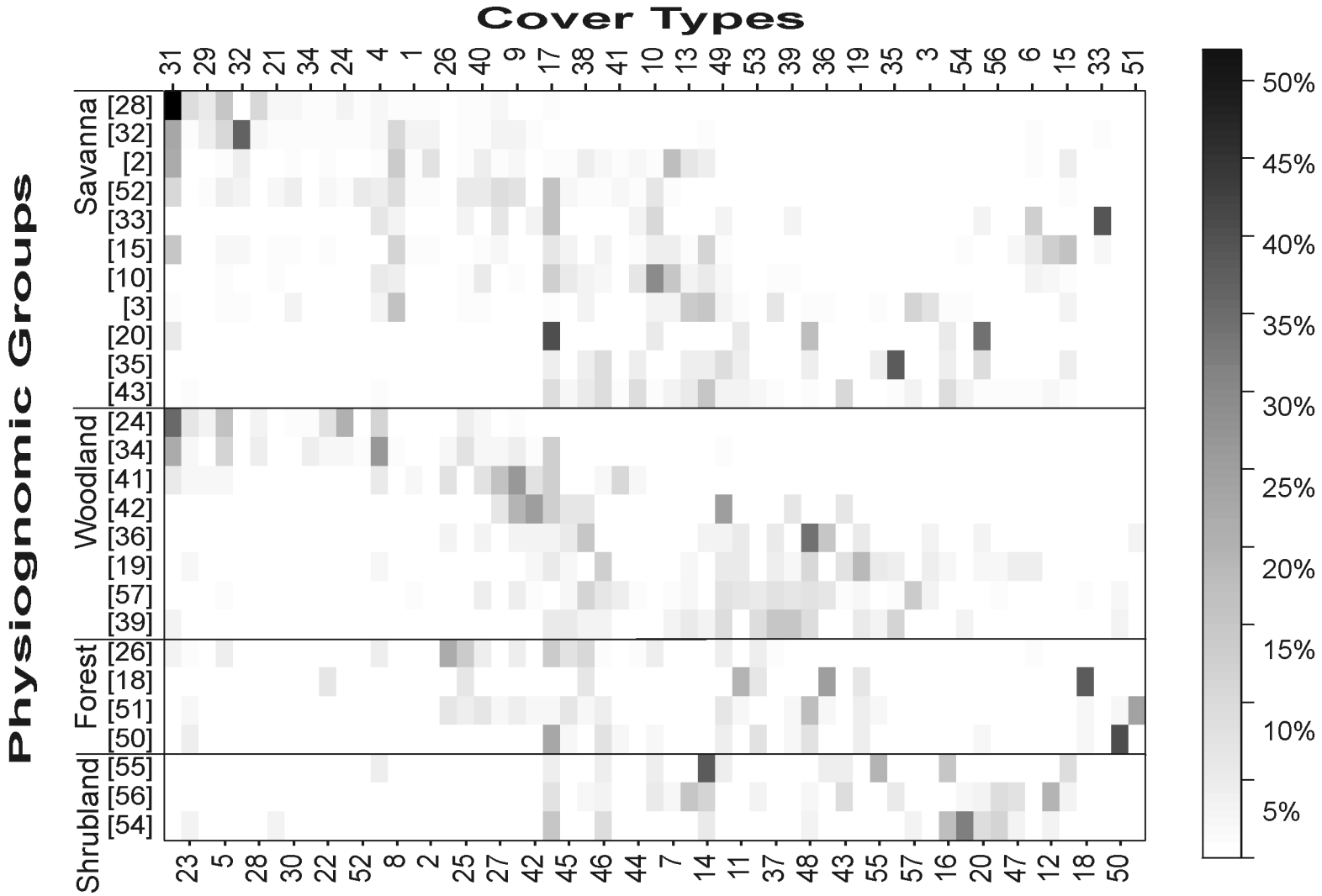


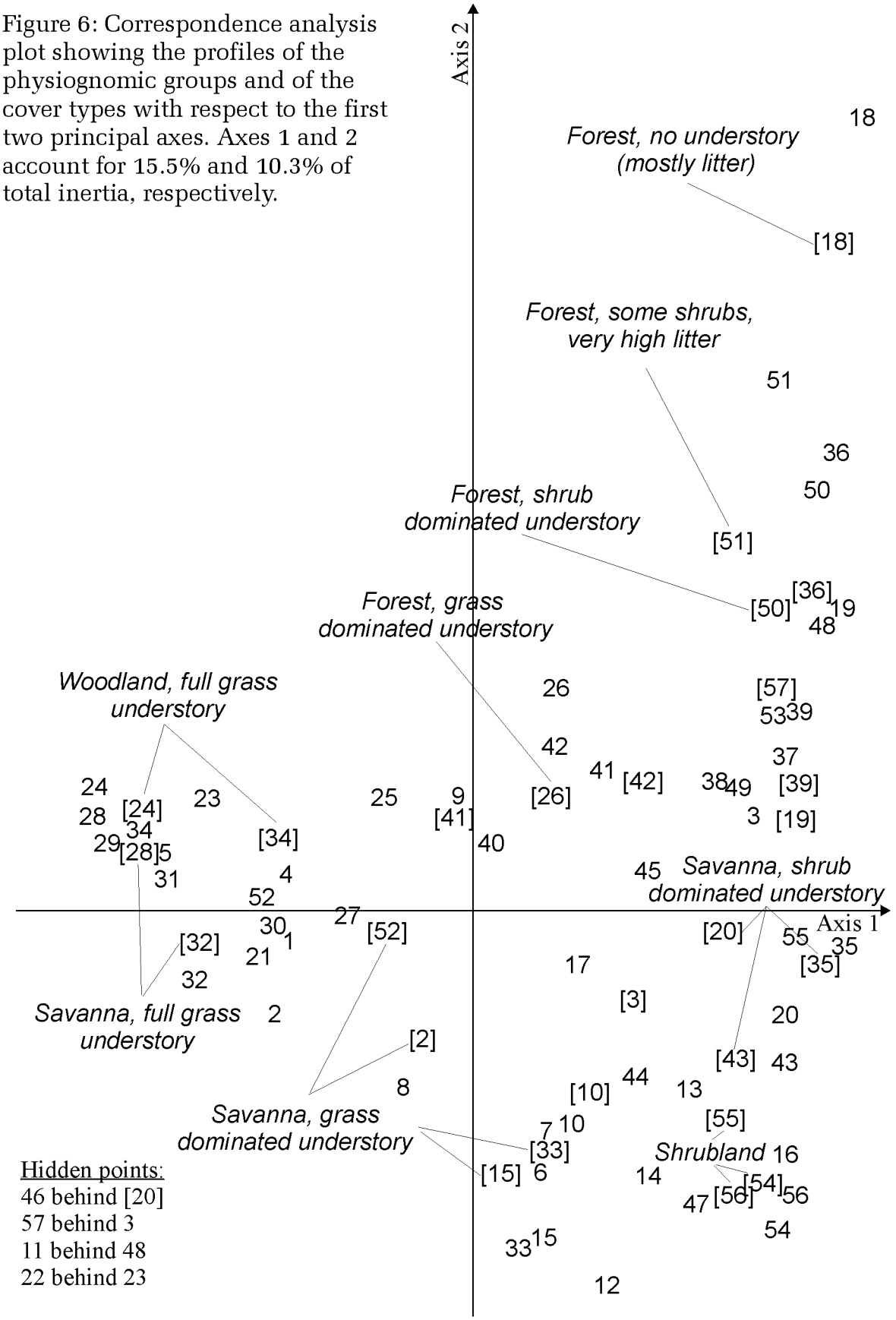
Figure 5: Density plot of the profiles of the physiognomic groups' composition in terms of cover types. Note: each row adds up to 100%.

Table 5: Physiognomic groups.

ID	*n*	*Description*
		Savanna
[28]	310	Full grass understory
[32]	160	Full grass understory, some understory oaks
[2]	81	Grass dominated understory, many understory oaks
[52]	99	Grass dominated understory, few shrubs
[33]	35	Grass dominated understory, much herb
[15]	183	Grass dominated understory, some shrubs and understory oaks
[10]	105	Grass – shrub dominated understory.
[3]	69	High litter cover, many understory trees.
[20]	20	Shrub dominated understory, high herb
[35]	24	Shrub dominated understory, some understory oaks, herb.
[43]	91	Shrub dominated understory, some understory oaks.
		Woodland
[24]	127	Full grass understory.
[34]	83	Full grass understory, few shrubs.
[41]	44	Grass dominated understory, few shrubs.
[42]	18	Some shrubs, low grass.
[36]	40	Some shrubs, no grass.
[19]	55	Some shrubs, understory oaks and herb.
[57]	105	Some shrubs and understory oaks.
[39]	39	Shrub dominated understory, some understory oaks.
		Forest
[26]	68	Grass dominated understory, some shrubs.
[18]	18	No understory (mostly litter).
[51]	48	Some shrubs, very high litter cover.
[50]	46	Shrub dominated understory.
		Shrubland
[55]	24	Many understory trees.
[56]	60	Many understory trees, some grass.
[54]	40	Mostly shrubs.

Note: the bracketed numbers are the groups' identifiers, they refer to the number in Table 1 of the cover type that was predicted by the decision tree in Fig. 1a.

Figure 6: Correspondence analysis plot showing the profiles of the physiognomic groups and of the cover types with respect to the first two principal axes. Axes 1 and 2 account for 15.5% and 10.3% of total inertia, respectively.



To validate the clustering into physiognomic groups, 1022 VTM plots of unknown hardwood cover type were classified into physiognomic groups using the decision tree of Fig. 2. The distribution of this validation data set among the physiognomic groups was very similar to that of the clustering data set (Fig. 7). The main difference between these two sets was that about 4% less plots fell within savanna groups and 4% more in the shrubland groups. The centroids of the 26 clusters of the validation set as differentiated by the three main variables can be found in Fig. 8. With the exception of group [20] that had a much lower shrub cover in the validation set, all groups 'centroids were consistent with those of the clustering set. To further investigate the correspondence between clusters from the clustering data set and those from the validation set, we computed species constancy for each set and for each physiognomic group in each set. The constancy of a species is the percentage of plots featuring that species in a group of plots. The validation set and clustering set had similar overall composition (Table 6), although with a lower incidence of *Q. Douglasii* and *Q. agrifolia*, and a higher incidence of *Q. Kelloggii* and *Q. chrysolepis*. Examination of species constancy by physiognomic group revealed a good overall correspondence between the clustering and validation sets (see example of group [32] in Table 6). Only group [20] had different species composition in the clustering and validation sets, which may be related to the fact that the number of plots involved in this group were quite small (20 and 12 plots respectively). Constancy of the main species for the combined 3014 plots is shown as a density plot in Fig. 9. This plot reveals some strong species-physiognomic group associations. For instance, *Q. Douglasii* has high constancy in savanna and woodland types with a grass dominated understory while *Q. agrifolia* and *Q. Kelloggii* have higher constancy in savanna and woodland types with a shrub dominated understory. *Q. agrifolia* has high constancy in forest types with high litter understory. *Q. Kelloggi* is associated with shrubland types. Some physiognomic groups have virtually no understory trees, some shrub species are found with more constancy in forest or shrubland groups than in savanna groups.

Table 6: Species constancy for plots in clustering and validation data sets.

Species	*Clustering set n=1992*	*Validation set n=1022*	*Clustering [32] n=160*	*Validation [32] n=85*
Overstory tree	%	%	%	%
Quercus Douglasii	53	46	90	86
Quercus Agrifolia	39	25	13	13
Pinus Sabiniana	24	24	22	53
Quercus Wislizenii	20	15	16	19
Quercus Lobata	19	11	12	11
Quercus Kelloggii	18	28	6	7
Aesculus californica	6	3	6	1
Arbutus Menziesii	5	3		
Quercus chrysolepis	4	17	1	
Understory trees				
Quercus Douglasii	19	20	74	80
Quercus Wislizenii	15	14	14	12
Quercus agrifolia	9	4	4	9
Quercus Kelloggii	6	12	1	
Aesculus californica	5	4	1	
Umbellularia californica	5	4	2	1
Quercus chrysolepis	3	11	1	
Shrubs				
Rhus diversiloba	35	23	14	13
Heteromeles arbutifolia	18	12	3	
Ceanothus cuneatus	16	15	8	7
Rhamnus californica	13	6		1
Rhamnus crocea	13	12	10	9
Artemisia californica	9	5	2	1
Arctostaphylos viscida	9	12	1	1
Lonicera subspicata	5	6	10	7
Arctostaphylos manzanita	5	4	2	
Diplacus auriantacus	5	2		
Holodiscus discolor	5	1		
Ceanothus integerrimus	4	9		1
Cercocarpus betuloides	4	8	2	
Chamaebatia foliosa	1	7		
Arctostaphylos glauca	3	4	1	6
Herbs				
Pteridium aquilinum	5	4		
Haplopappus linearifolius	2	4	5	6
Grasses	79	76	100	100
Ground				
Litter	49	47	1	4
Bare soil	13	21	3	7
Rock	12	13	18	7

Note: only species with a constancy of 5% or more in one of the columns are shown.

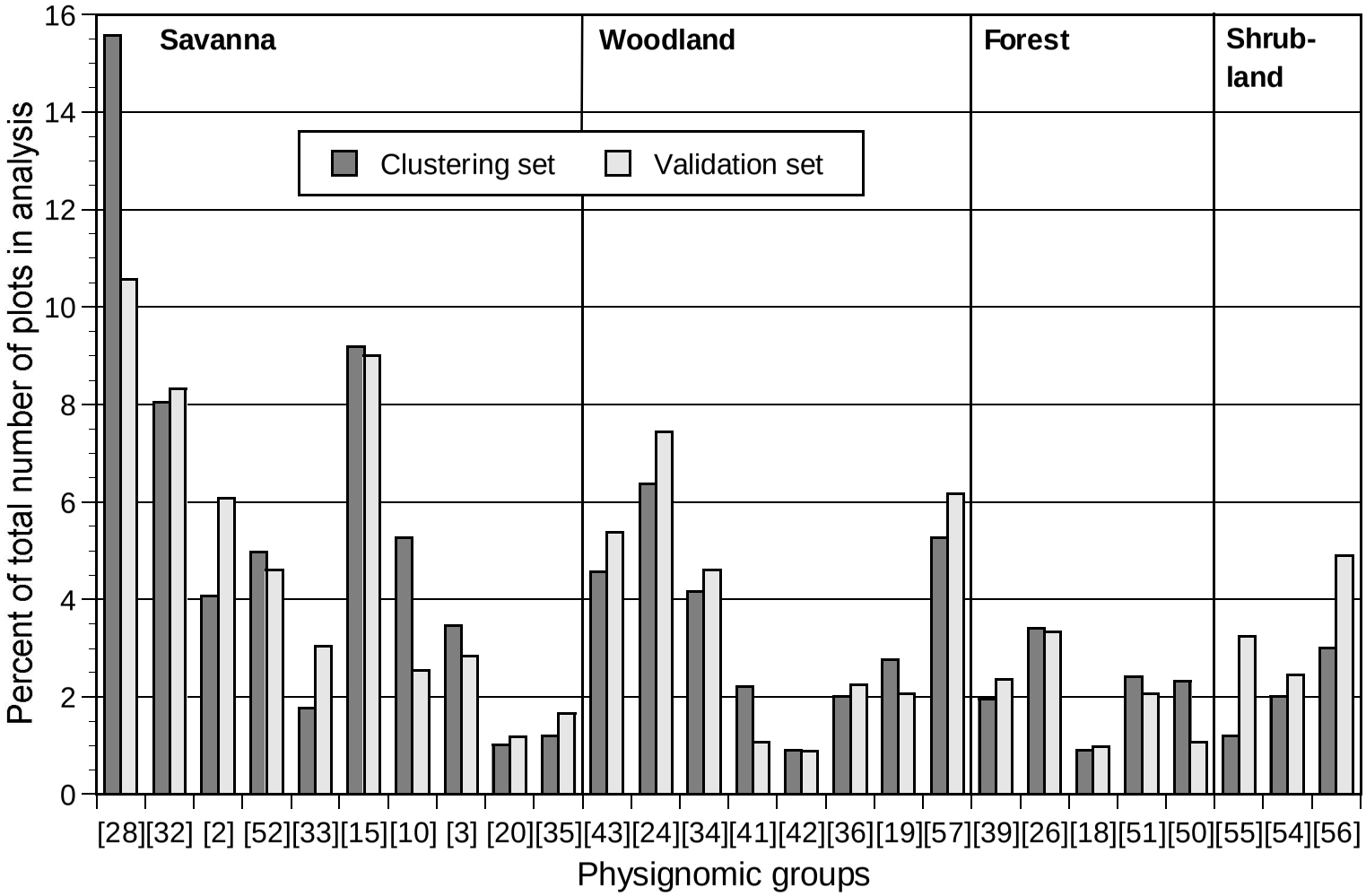


Figure 7: Distribution of the clustering and validation data sets into physiognomic groups.

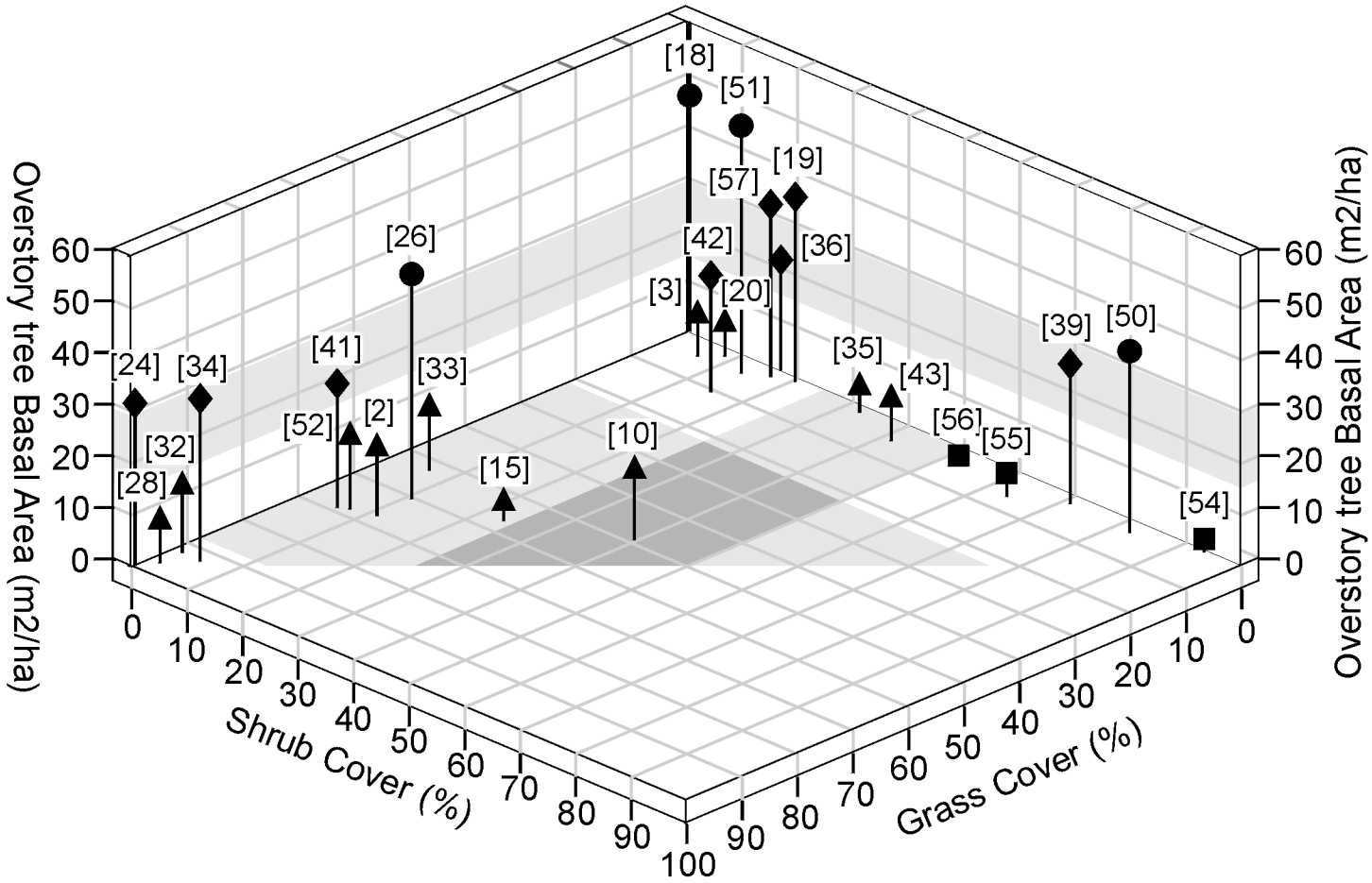


Figure 8: Cluster centroids of physiognomic groups for the validation data set. The shading shows variable ranges determined by the main split in the corresponding decision tree. Markers correspond to the main vegetation types of Table 5: triangle = savanna, diamond = woodland, circle = forest, square = shrubland. Cluster identifiers are shown over the markers.

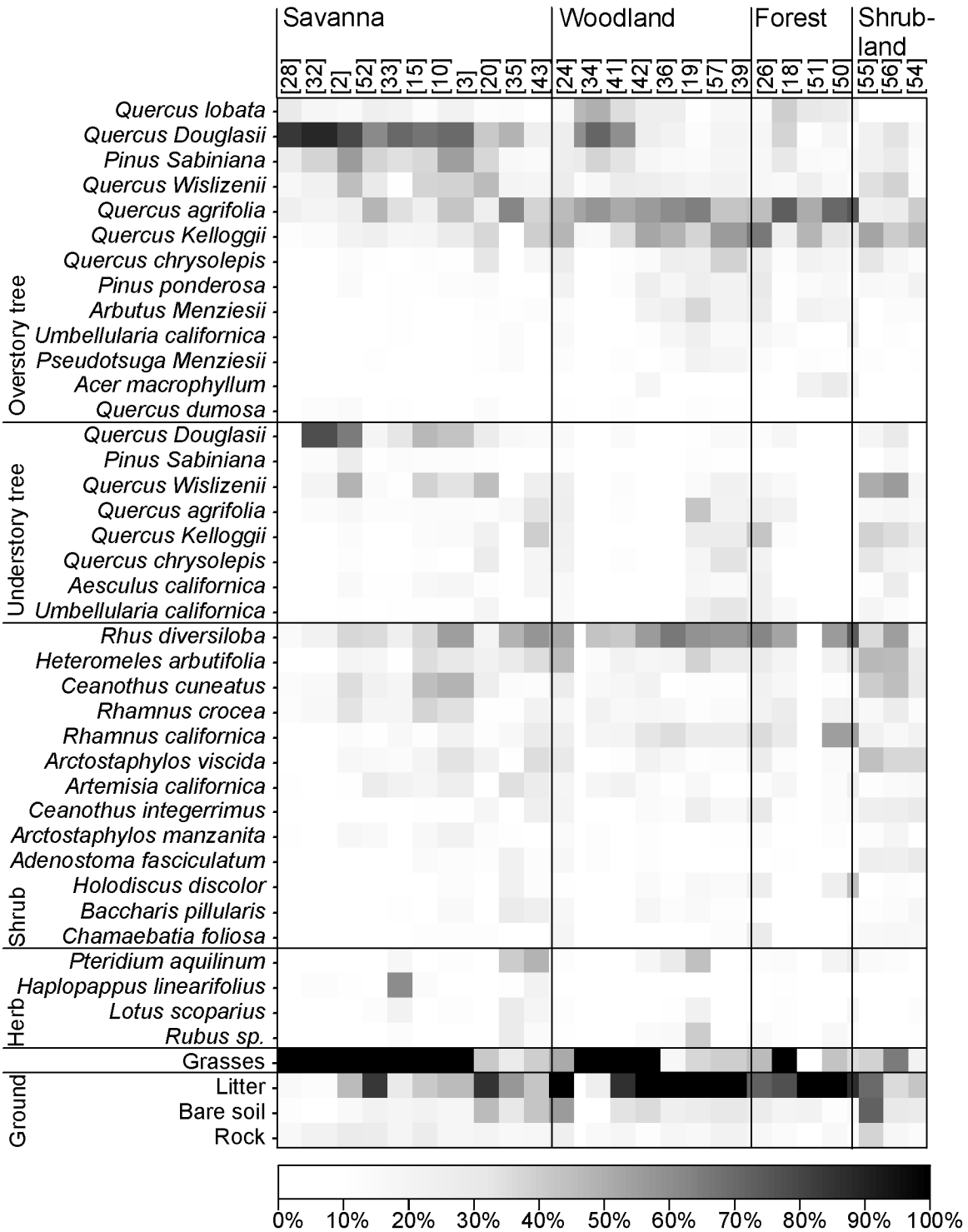


Figure 9: Species constancy for clustering and validation sets combined (n=3014). Note: only species with a constancy of 15% or more in at least one of the physiognomic groups are included.

Grouping of physiognomic groups into abiotic domains.

This third analysis is another case of supervised clustering. We used CART to predict the physiognomic groups (Table 5) from abiotic factors reflecting geographical position, topography, soil and climate (Table 2). The diversity of such relationships at the scale of the entire state of California resulted in too much complexity for a single analysis. The analysis presented here was, restricted to the region east and north of the Central Valley. This region (Fig. 1) comprises 1177 of the 3014 plots classified into physiognomic groups in the second analysis. The decision tree we have retained is presented in Fig. 10. The first split on average annual precipitation (at 607 mm / year) separates the driest part of the Sierra Nevada foothills and adjacent valley floor plots from the others. Two more splits involved precipitation (at 1117 and 1831 mm / year) delineating a mesic, a wet, and a high precipitation zones in the region considered (Map 1 and 2). Further splits in the mesic zone were as follow. A split on mean minimum temperature in January (at 2.43° C) separates out sites with mild winters in the Sierra Nevada foothills. Among sites with colder winters a split on mean temperature range in January (at 12.1°C) separates sites with greater winter temperature fluctuations. A last split is made on mean maximum temperature in July (at 35.9°C) isolating sites with hot summers, mostly at the northern extremity of the Central Valley. Within the wet zone, a single further split was made on distance to the coast (at 172 km) separating sites north of the Central Valley, in an area west of Shasta lake from a larger group of sites along the Sierra Nevada. The high precipitation zone comprises a small group of sites at higher altitudes in the northern Sierra Nevada.

The centroids of the 8 clusters corresponding to the leaf nodes of the decision tree in Fig. 10 can be found in Fig. 11. In this view the clusters are differentiated by the three main variables (precipitation, mean minimum temperature in January and mean maximum temperature in July) involved in the decision tree. The January minimum temperature (JAMI) was chosen for this graph because it was indicated by CART as a surrogate variable for the splits on coast distance (COASTDST) at a level of 1.15°C and on January temperature range (JARAN) at a level of minus 0.8°C. A split on a surrogate variable is similar to the best split and is used by CART for cases with missing data (Breiman et al. 1984). As expected, the centroids of the clusters in Fig.11 show that JAMI can discriminate between B and (D and C) substituting for JARAN; and also between F and G, substituting for COASTDST.

In this decision tree, two of the leaf nodes (B and D) have been allocated by CART to the same physiognomic group (group [15]). If this was a supervised classification analysis, these two nodes could be readily aggregated since they predict the same outcome. However, in a supervised clustering analysis we have seen that we must first decide if the proportions of physiognomic groups found in the two nodes are similar enough to justify merging the two clusters. The density plot (Fig. 12) of the profiles of the abiotic domains 'composition in terms of physiognomic groups provides a visual way of comparing clusters B and D. Physiognomic group [15] is the archetype of both clusters, and they both have a majority of their plots coming from savanna groups, mostly with grass-dominated or full-grass understory. Both clusters include very few woodlands and almost no forest groups (about 5% forest for B and 1% for D). They have similar proportions coming from the shrubland groups and from savanna with shrub dominated understory (group [43]). The plot of the correspondence analysis of abiotic domains and physiognomic groups (Fig.13) also shows a strong similarity between B and D. We decided to merge cluster B and D into a single domain called B&D. Further examination of Fig. 12 and 13 reveals a sequence of clusters from domain A regrouping mostly savanna and woodland plots with a full grass understory, to domain H composed principally of woodland and forest with a shrub dominated understory. Following domain A in the sequence are B&D, C and E, which are composed principally of savanna with an understory combining grass and shrubs. Then come clusters F and G, gathering mostly savanna and woodland plots with an understory dominated by shrubs and shrubland plots. Fig 11 shows that this sequence from full grass savanna and woodland to shrubby woodland and forest corresponds to a sequence of zones with increasing precipitation levels. Vegetation states domains based on abiotic factors are summarized in Table 7.

The next step in our assessment of the "goodness" of our clustering into abiotic domains was to consider the overall species composition of the plots in these domains (Fig. 14). Here one can find a sequence of shifts in species constancy from domain A to domain H. In domain A to E, *Q. Douglasii, Pinus Sabiniana* and *Q. Wislizenii* are the most common species in the overstory. However, while *Q. Douglasii* is often the only tree species in domain A, in domain C and E *Pinus Sabiniana* and *Q. Wislizenii* are often co-occurring with it. These three species also co-occur often in domain B&D, but here *Q. Wislizenii* also replaces *Q. Douglasii,* and *Q. Kelloggii* is present in about a third of the plots. In domain A, incidence of

trees in the understory is lower than in B&D, C and especially E. Species composition in the understory tree layer reflects that of the overstory, although with a much lower incidence of *Pinus Sabiniana*. Incidence of shrub species is very low in domain A, but shrubs are common in domain B&D, C and E. These four domains are characterized by very low incidence or absence of species in the herb layer, very high incidence of grasses and low to moderate incidence of litter with the exception of domain E where litter is found in every plot. Species composition in domains F, G and H is quite different from that in the previous four domains. Here *Q. Kelloggii* is present in nearly every plot, and in most cases it is the dominant species. Three other species of trees are often present in these domains: *Q. chrysolepis, Pinus ponderosa* and *Pseudotsuga Menziesii*, with constancy increasing from low in domain F to high in domain H. Here also, understory tree layer composition reflects that of the overstory, although with a lower incidence of all species in domain H. Incidence of shrub species is high overall, with the presence of a series of species that are not found in domains A through E. Herb species are much more common here, especially in domain F. Grasses as a group are a component of the understory in only a quarter of the plots in domains F and G and 10% of those of domain H. Litter is often a component in F and G, present in every plot in domain H.

Table 7: Vegetation states domains based on abiotic factors (region east and north of the Central Valley).

ID	*n*	*Position*	*Description*
A	296	Driest part of the Sierra Nevada foothills and adjacent valley floor, Tehachapi mountains. (ppt ≤ 600 mm / year).	Full grass understory, savanna and woodland. Shrubs and understory oaks are rare.
B&D	526	Mesic zone of the Sierra Nevada foothills (600 ≤ ppt ≤ 1100 mm / year).	Grass dominated understory, savanna and woodland. Shrubs and understory oaks are common.
C	33	Mesic zone with hot summers. Mostly at the north end of the Central Valley.	Shrub-Grass dominated understory, savanna.
E	95	Mesic zone with mild winters. Sierra Nevada foothills, mostly by Folsom lake, American and Cosumnes river.	Savanna, grass dominated understory, many understory oaks.
F	51	Wet zone (1100 ≤ ppt ≤ 1800 mm / year) with mild winters. East Shasta lake area.	Savanna and woodland. Shrub dominated understory.
G	144	Wet zone with cold winters. Sierra Nevada and Southern Cascades.	Shrubland, savanna and woodland. Shrub dominated understory.
H	32	High precipitation zone of the Sierra Nevada (> 1800 mm / year). Cold winters, at the upper limit of oak range.	Woodland and forest. Shrubs and understory trees are common, understory often mostly litter.

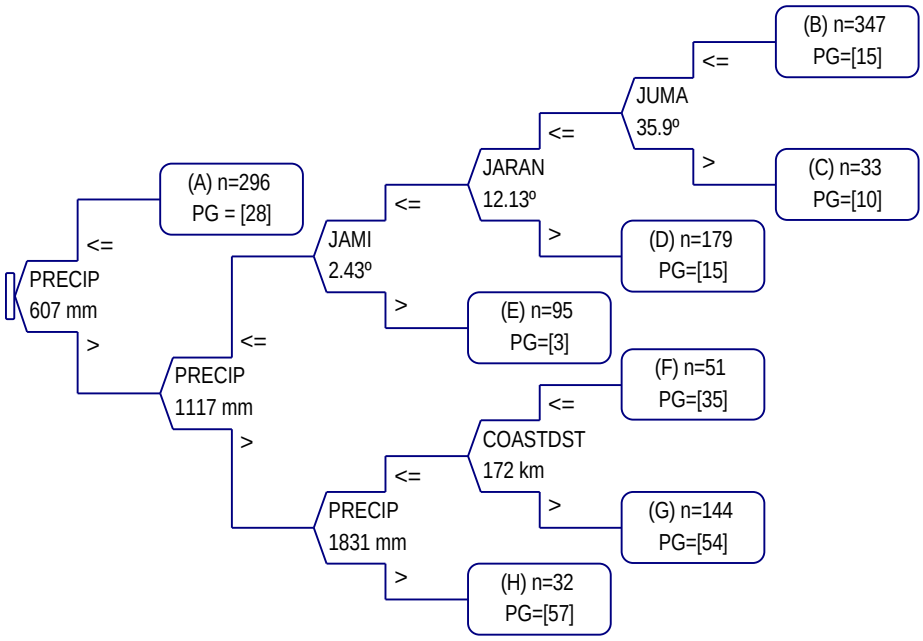


Figure 10: Decision tree of the clustering of plots into abiotic domains. The root node of the tree is on the left and the branching nodes are marked with the corresponding split variable and split value. The leaf nodes are marked with (1) the cluster identifier, (2) number of plots in the cluster, (3) the physiognomic group predicted for the node.

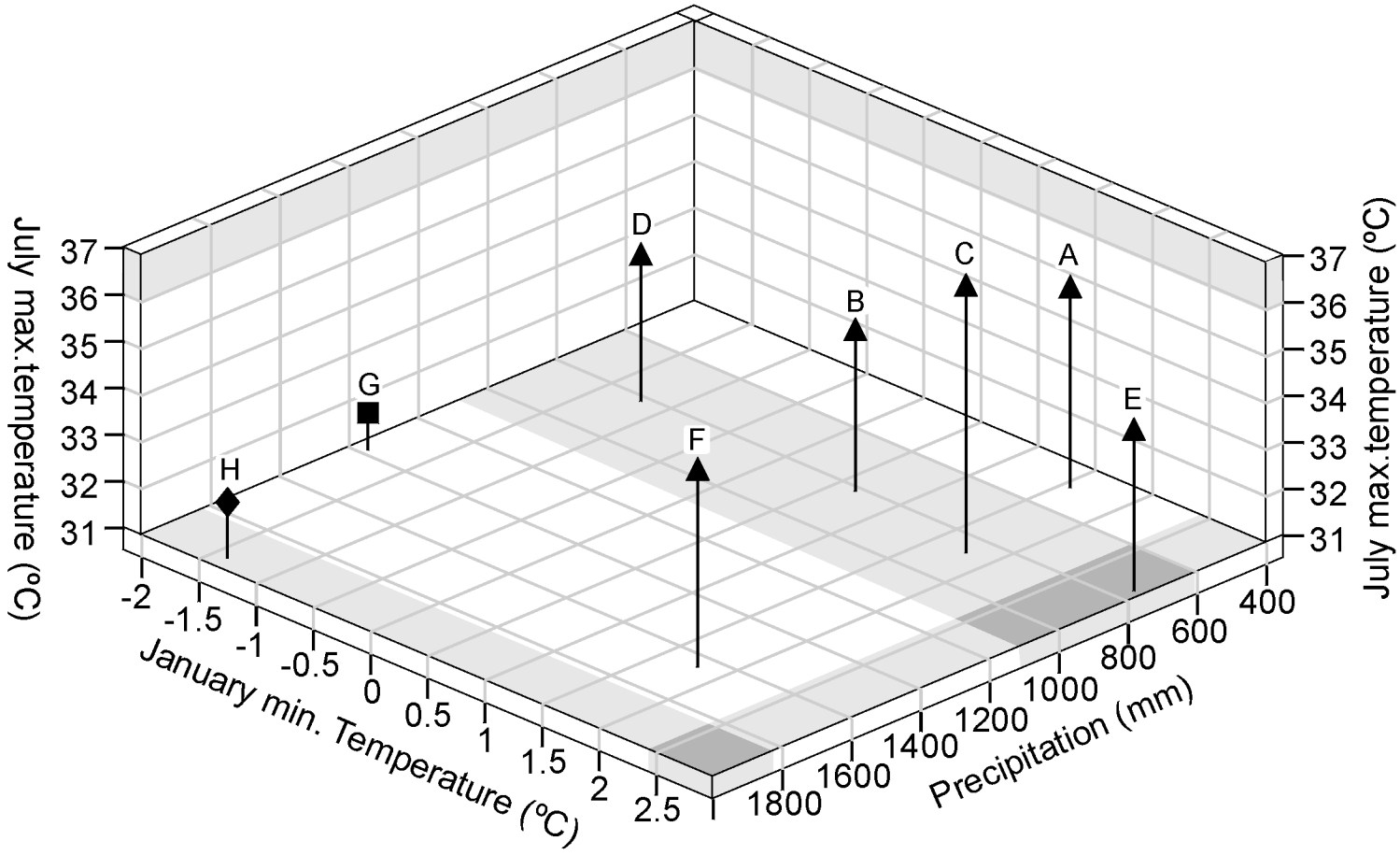


Figure 11: Cluster centroids of abiotic domains. The shading shows variable ranges determined by the main split in the corresponding decision tree. Markers correspond to the main vegetation types of Table 5: triangle=savanna, diamond=woodland, circle=forest, square=shrubland. Cluster identifiers are shown over the markers.

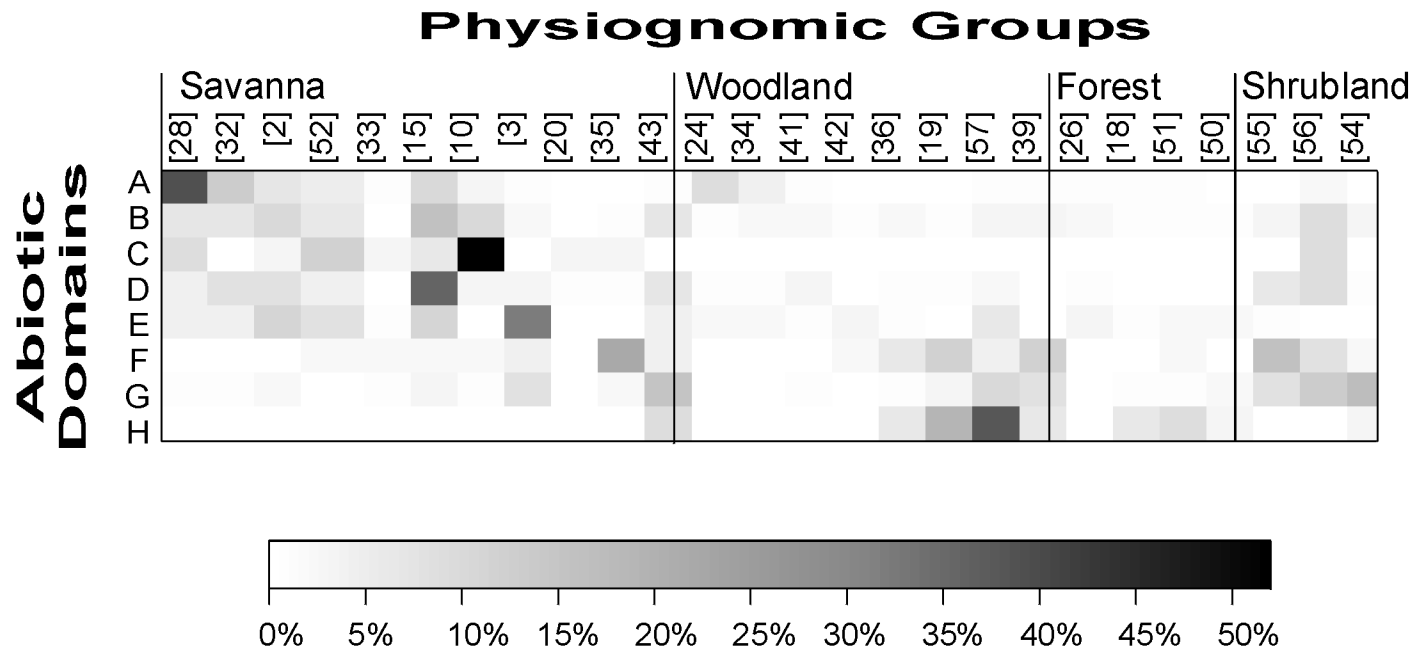


Figure 12: Density plot of the profiles of the abiotic domains' composition in terms of physiognomic groups. Note: each row adds up to 100%.

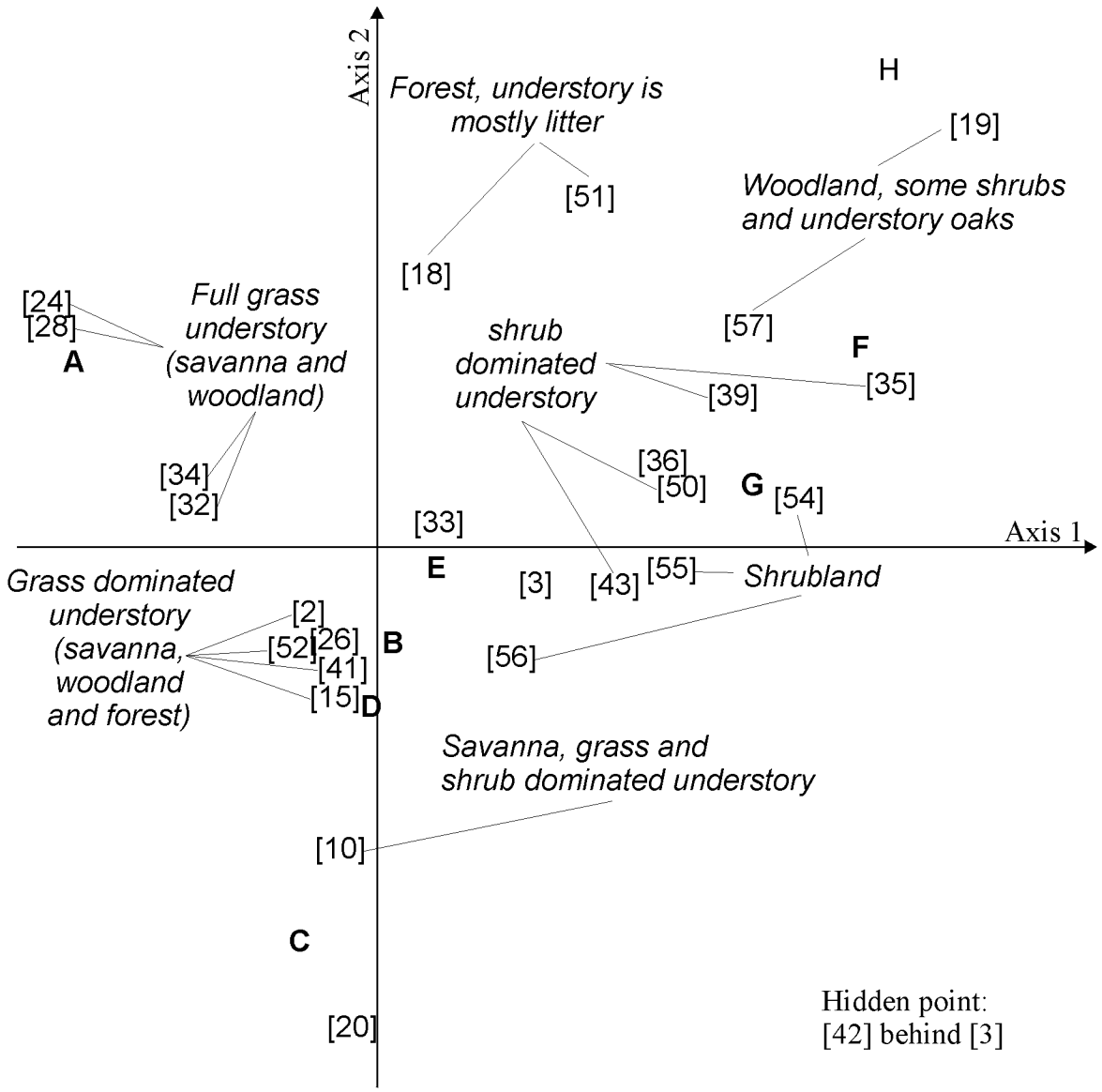


Figure 13: **Correspondence analysis plot showing the profiles of the abiotic domains and physiognomic groups with respect to the first two principal axes. Axes 1 and 2 account for 36.1% and 17.3% of total inertia, respectively.**

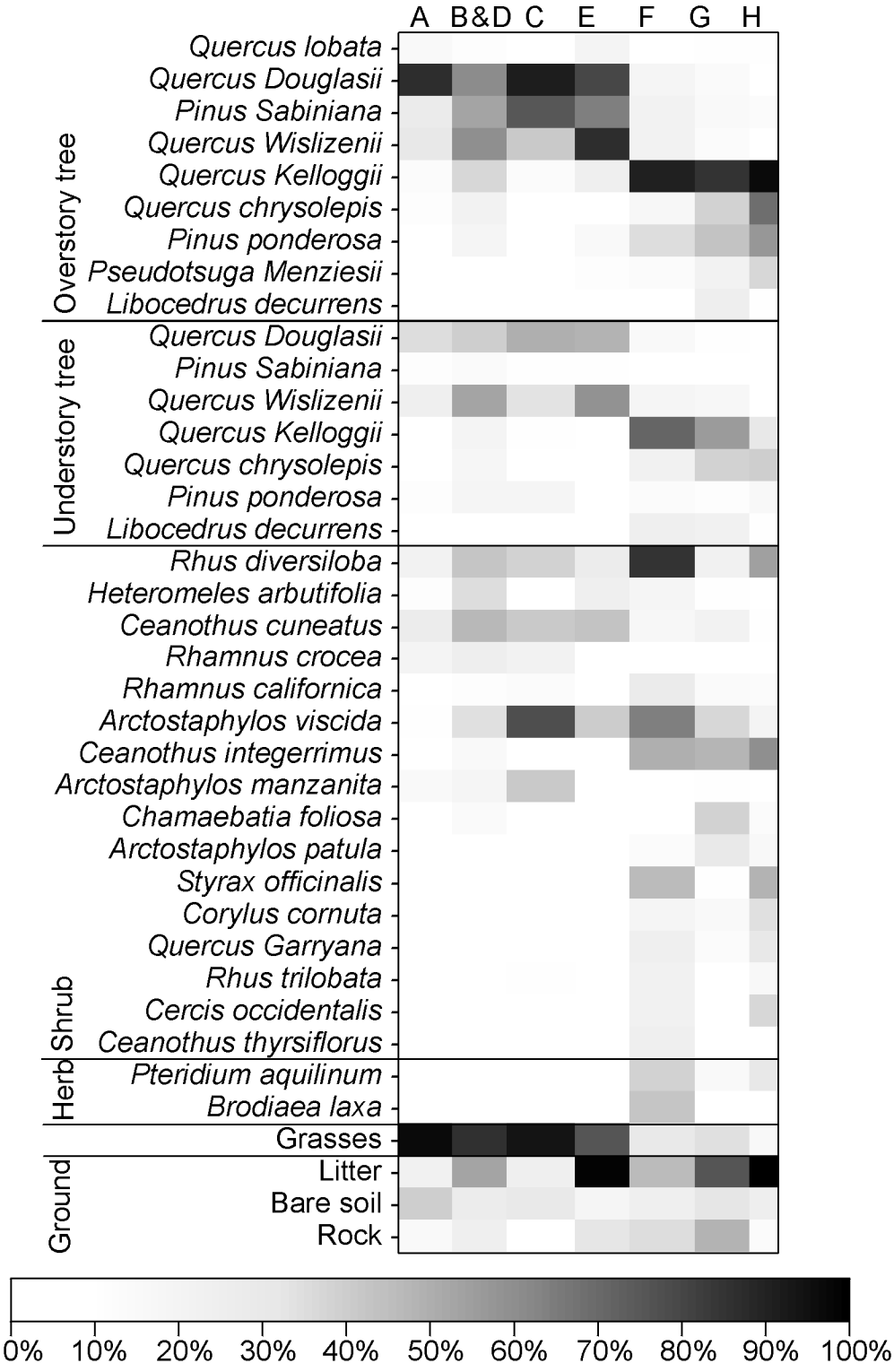


Figure 14: **Species constancy for abiotic domains north and east of the Central Valley (1177 plots). Note: only species with a constancy of 15% or more in at least one of the physiognomic groups are included.**

To validate the clustering into abiotic domains we used a different data set: California Department of Forestry ś GIS database of California ś hardwood rangelands (Fig. 1). Since this GIS database was developed from aerial photographs and Landsat Thematic Mapper images, its coverage is much more extensive than that of the VTM plots but the vegetation data it contains is much less precise. The information about each polygon is limited to a classification based on wildlife habitat relationships (WHR, Mayer and Laudenslayer, 1988): percent canopy closure (5 classes) and WHR hardwood cover types (5 classes). This classification system characterizes only the overstory tree layer because information about trees, shrubs, herbs and grass in the understory is either hidden or too difficult to interpret from the sky. Even in the overstory, determination of tree species by remote sensing is often difficult, with some confusions possible (Pacific Meridian Resources, 1994). Given the limitations of this system for identifying specific species assemblages (Table 8) we restricted our validation to precipitation zones (Fig. 11) in the region north and east of the Central Valley.

Table 8: WHR cover types with their primary and associate species composition (from Pillsbury et al. 1991) for north and east of Central Valley region.

WHR type	***Blue oak woodland (BOW)***	***Blue oak / Foothill pine (BOFP)***	***Valley oak woodland (VOW)***	***Montane Hardwood mix (MH)***
Primary species	*Quercus Douglasii*	*Q. Douglasii* *Pinus Sabiniana*	*Q. lobata*	*Q. Kelloggii* *Q. Wislizenii* *Q. chrysolepis* *- Arbutus Menziesii* *- Lithocarpus densiflorus*
Associated species	*Q. Wislizenii* *Q. agrifolia* *Q. chrysolepis* *Q. lobata* *Aesculus californica*	*Q. Wislizenii* *Q. agrifolia* *Q. chrysolepis*	*Q. Wislizenii* *Q. Douglasii* *Q. agrifolia* *Q. chrysolepis* *Pinus Sabiniana* *Q. Garryana*	*Q. Garryana* *Q. agrifolia* *Q. lobata* *Q. Douglasii* *Pinus Sabiniana* *Pinus ponderosa*

The results of our GIS analysis are presented in Fig. 15. The total area of hardwood rangelands in the region was 1,692,950 ha: 638,900 ha located in the driest zone (< 600 mm/year); 973,425 ha in the mesic zone (600 –1100 mm);

80,625 ha in the wet zone (1100 –1800 mm) and only 425 ha in the high precipitation zone (>1800 mm). This last zone was left out of the analysis because it represented little area compared to the others. The first graph in Fig. 15 presents the distribution of the 5 classes of hardwood canopy closure within each of the three main precipitation zones. It shows that percent canopy closure increases with amount of yearly precipitation. This is in agreement with the trend we found from a predominance of savanna and woodland in domain A to mostly woodland and forest in domain H (Fig. 12 and 13). The second graph in Fig. 15 presents the distribution of the WHR hardwood cover types (Table 8) within the three main precipitation zones. Most of the hardwood rangelands in the driest zone are blue oak woodlands (BOW) types. In the mesic zone blue oaks woodlands (BOW), blue oak and foothill pine (BOFP) and mountain hardwood mix (MH) types are about equally represented. In the wet zone, mountain hardwoods constitute most of the hardwood rangelands. This trend confirms that found in our third analysis (Fig. 14): plots in domain A (in the driest zone) are mostly dominated by *Q. Douglasii* alone; plots in domains B&D, C and E (in the mesic zone) are populated by *Q. Douglasii, Pinus Sabiniana* and *Q. Wislizenii* (and with lesser constancy *Q. Kelloggii*) occurring together or alternatively; plots in domains E and F (in the wet zone) are mostly dominated by *Q. Kelloggii,* with *Q. chrysolepis* and *Pinus ponderosa* having the next highest incidence.

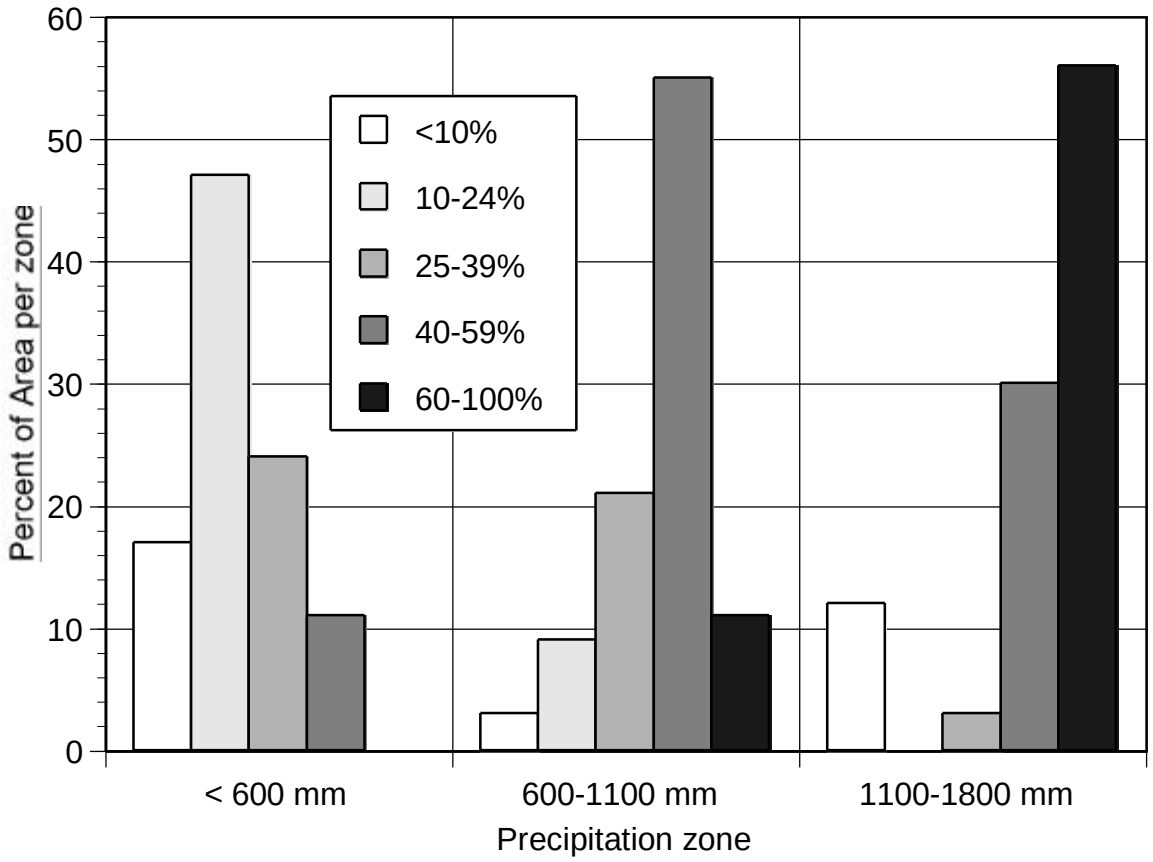


Figure 15 (a): Distribution of canopy closure classes in zones of increasing yearly precipitation (see Fig. 11) in the region north and east of the Central Valley.

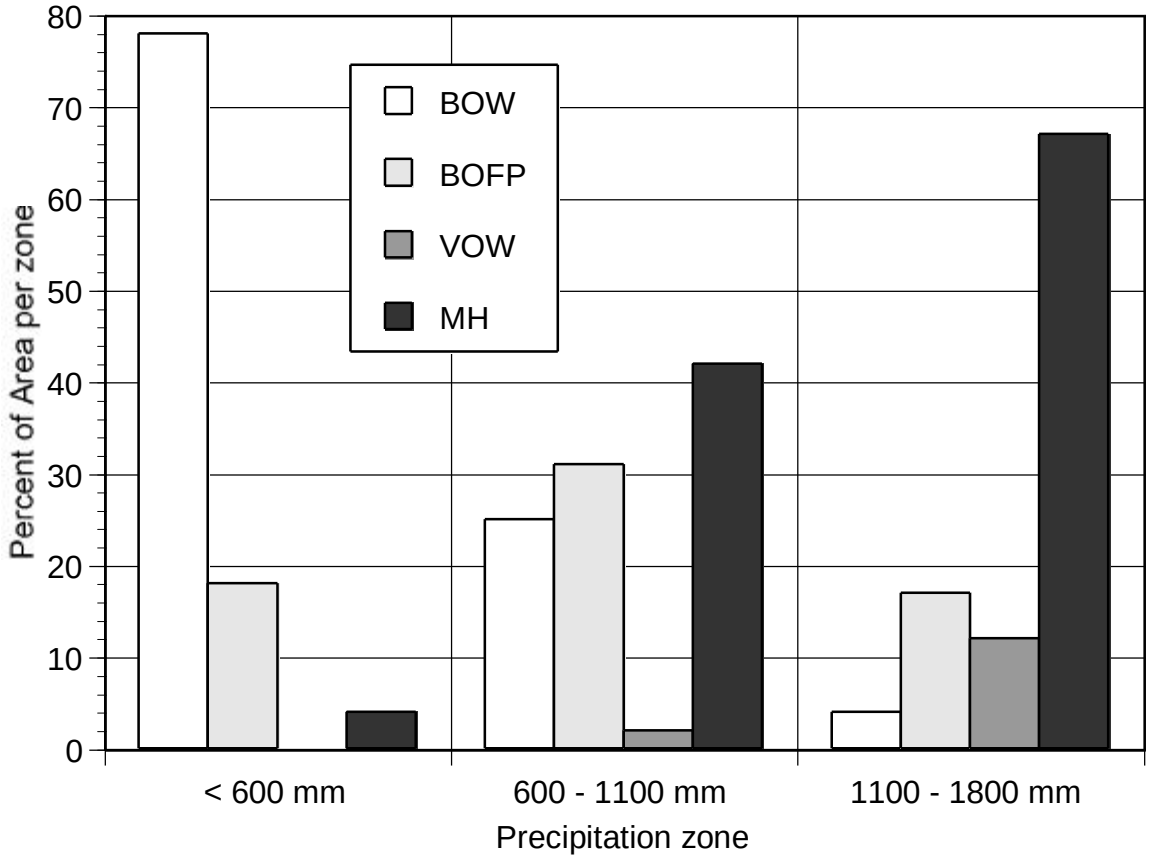


Figure 15 (b): Distribution of WHR vegetation types (see text for classes) in zones of increasing yearly precipitation (see Fig. 11) in the region north and east of the Central Valley.

Example of a State-and-transition model.

The results of the analyses above were presented to a group of hardwood rangeland experts. A full day workshop produced a tentative S&T model for domain A. The diagram of the states and transitions can be found in Fig. 16. The physiognomic groups [28], [32], [15], [24], [2], [52] and [34] which characterize 87% of the 296 plots in domain A served as a basis for the vegetation states of Fig. 16. Two vegetation states (I and II) were added to those derived from the physiognomic groups and two groups [24] and [34] were lumped into a single vegetation state. The catalogs of states and transitions is presented in Appendix 3.

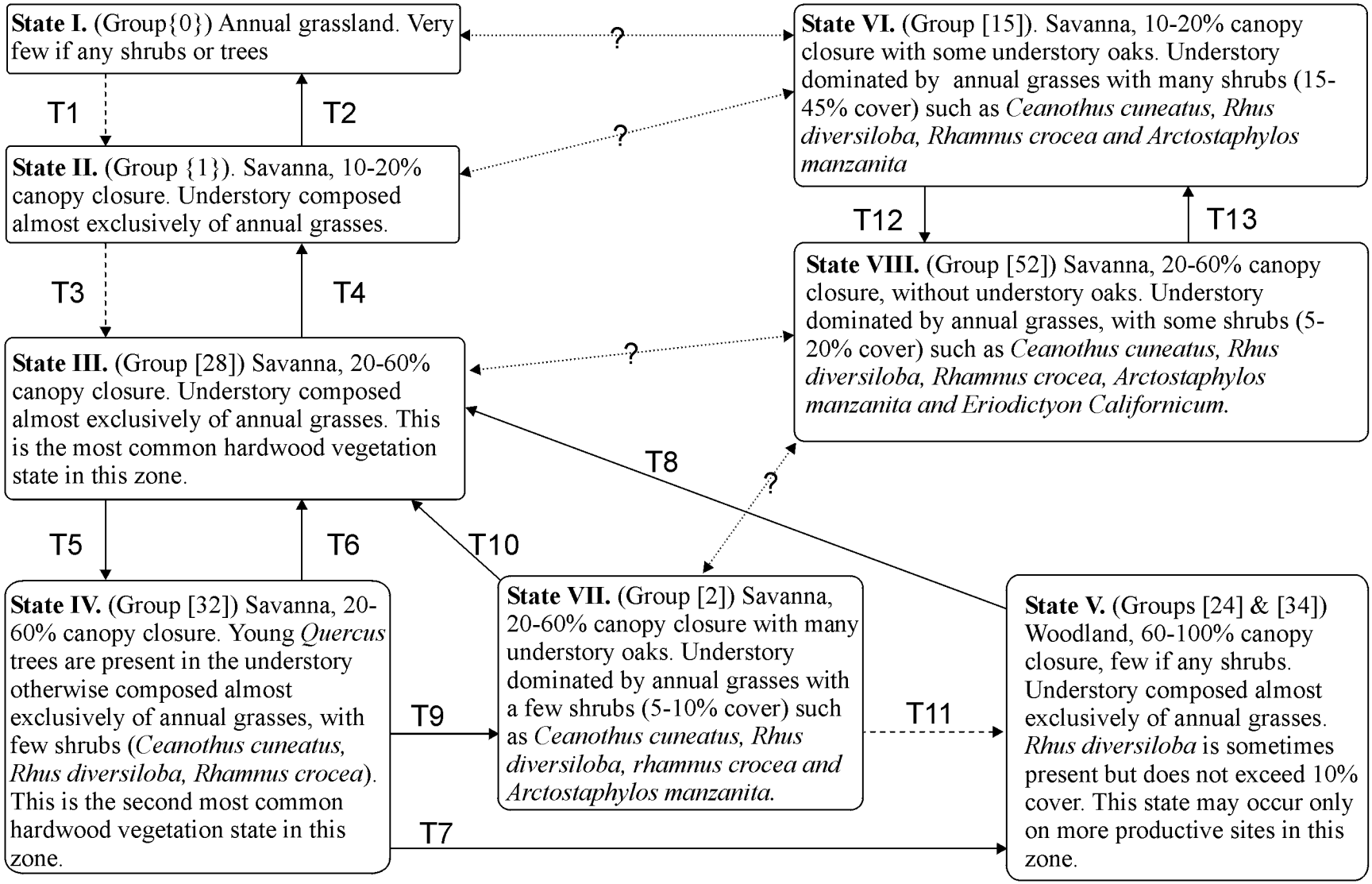


Figure 16: State-and-transition model for domain A.

Discussion.

Cluster assessment and ecological insight.

A quantitative key to the cover types.

The first analysis shows how a conceptual classifier such as CART can help interpret a given classification of vegetation. CART can help build keys to the cover types that are more quantitative because of the way it uses computer power to search for the best split values on key variables. The use of quantitative keys is less subject to users 'interpretation and thus permits a more consistent determination of the cover type of new cases. Quantitative keys are also desirable for their straightforward computer implementation. Importantly, through cross-validation CART also helped eliminate spurious complexity and avoid overfitting the data that served to construct the classification. This helps design keys that classify new cases with better accuracy and efficiency. We have also shown how conceptual classifiers can help derive concepts for clusters that may be known only in terms of the list of cases they contain. Mirkin (1996) called it finding a "theoretical", intensional cluster structure to describe an "empirical", extensional one that is given. In sum, the first analysis showed how a supervised classification method can be used to supplement conventional cluster analysis.

Better descriptors of vegetation dynamics.

In our second analysis we have used CART in an unconventional manner: to do a supervised clustering analysis. What did we achieve through supervised clustering that we could not have done using traditional unsupervised clustering in this case? Here, we have used conceptual clustering as a tool to simplify a given classification while keeping some of the information contained in it. More precisely, we extracted from a given classification the structures that relate to a subset of the object attributes. When we decided to use a physiognomic rather than a floristic grouping to help delineate vegetation states, we made attempts to produce physiognomic groups with unsupervised clustering methods and ordination. The problem was that no "natural" groupings could be found on life-form abundance data alone. Because there was no clear discontinuities in the relative abundance of the different life forms, the clusters that were produced by conventional means were quite arbitrary. In contrast, the physiognomic groups we obtained through supervised clustering were much more satisfactory. They

correspond well to the common classification of hardwood rangelands into savannah, woodland and forest. Beyond these distinctions based on the overstory, our physiognomic groups also delineate combinations of life forms abundance that have important management implications. For instance, grass cover in the understory conditions the value for grazing livestock, while the amount of overstory tree cover affects the productivity of the grass-forb understory. Shrubs need to reach a certain level of cover before they constitute a ladder-fuel that may allow ground fires to reach into the tree crown. Yet, crown fires will kill many trees only if overstory tree canopy closure is high enough.

To be useful in this case, the supervised clustering analysis must provide information that goes beyond extracting basic structural distinctions from the cover type classification. Otherwise, it would be more straightforward to directly define physiognomic groups that reflect structural distinctions and other combinations of life form abundance with management implications. The associations we have found between our physiognomic groups and some groups of species (Fig. 9) indicate that indeed new information was generated. Moreover, the validation of the physiognomic groups showed that the decision tree could be used to classify additional plots into groups that were consistent in terms of species composition (constancy) with those derived through the analysis. With supervised clustering, our physiognomic groups were derived from life form abundance but under supervision of a floristic classification. Thus, the patterns in physiognomy emphasized by the analysis are those most related to patterns in relative species abundance.

The lower incidence of *Q. Douglasii* and *Q. agrifolia*, and higher incidence of *Q. Kelloggii* and *Q. chrysolepis* in the validation set (Table 6) are consistent with a shift in number of plots from full grass savanna to shrubland groups (Fig. 7). *Q. Douglasii* is known for its adaptation to the most xeric part of the hardwood rangelands (Evett, 1994) and it is strongly associated with savanna groups with a grass-dominated understory. On the contrary, *Q. Kelloggii* is known to occupy the most mesic parts and it is strongly associated with woodland and forest with a shrub dominated understory, and the shrubland groups. *Q. chrysolepis* is the most widely distributed oak in California, and it assumes various growth forms in various conditions (Pavlick et al. 1991). We found it mostly associated with a shrub-dominated understory in woodlands and forest and with shrublands. *Q. agrifolia* is intermediate in drought resistance, mostly confined to zones of oceanic

influence. It is associated with shrubby savanna and especially with woodland and forest, with either grass, shrub or litter dominated understory. *Q. Wislizenii* exhibit intermediate drought resistance also. This makes it adapted to zones not quite as xeric as *Q. Douglasii* and it is found associated with more shrubs. Although found areas with low precipitation and high evaporative demand, *Q. lobata* does not have the drought adaptations of *Q. Douglasii*. It is thus restricted to sites where its roots can reach a water table. *Q. lobata* is not well represented in this study because most of its valley floor and riparian habitat had been already converted to agriculture when the VTM survey was conducted. Here, *Q. lobata* was associated mostly with woodland and forest with a grass dominated understory. These associations are only tendencies however, because physiognomy is the result of both site potential and a history of natural or management induced disturbance.

In the end, a "good" set of clusters must satisfy the purpose that motivated the analysis. This is the case here since we achieved a reclassification of cover types into groups that are more appropriate to describe vegetation dynamics (Fig. 5). For instance, *Quercus* / grass cover types (e.g. types 31, 23, 29, 5, 32, 28, 21, 30, etc in Table 1) were reorganized into savanna and woodland (e.g. physiognomic group [32] vs. [24] in Table 5). This first distinction is necessary to describe a decrease in overstory tree density resulting from a wood harvest or an increase in density from the maturation of a cohort of saplings. This distinction, which was impossible in the cover type system, has implications in terms of the effect of canopy closure on forage production among others. At the same time, the plots with *Quercus* / grass cover types were reshuffled into groups with full-grass or grass-dominated understory (e.g. group [28] vs. [52]), which reflects the presence or absence of shrubs in the understory. Concurrently, *Quercus* / grass cover types were also reorganized into groups having some understory oaks and those without (e.g. group [32] vs. [28]) which reflects the presence or absence of regeneration, a subject of particular concern in hardwood rangelands. Further examination of Fig. 5 shows how other kinds of cover types (e.g. *Quercus* / shrub sp. / grass; *Quercus* / shrub spp.; *Quercus* – tree sp.; etc) have been similarly reorganized into groups useful to describe vegetation dynamics.

Arriving at a satisfactory set of clusters is not an automatic process. A preliminary analysis, conducted with CART assuming priors equal to the proportions of each cover type in the data, led to a much less satisfactory set of

clusters. This set, with 18 physiognomic groups derived from a decision tree with 22 leaf nodes, failed to reflect even the basic distinction into savanna, woodland and forest. This is not surprising given the choice of priors: cover types with few plots did not have a chance to serve as an archetype for a physiognomic group. An extreme case is cover type 18 (*Q. agrifolia*) which comprises only 9 of the 2038 plots classified by Allen et al. This cover type is structurally unique in having a very high basal area of trees and an understory with virtually no tree, shrub, herb or grass cover (only litter). When we re-run the analysis with equal priors for all cover types, this cover type served as an archetype for physiognomic group [18], regrouping 18 plots out of 1992 (Fig. 5) and marking one of the extremes of the range of physiognomies found in hardwood rangelands (Figs. 3 and 6). Further adjustments were necessary to arrive to the physiognomic groups presented here. In particular, small adjustments to the final size of the pruned tree allowed distinguishing between full-grass savannas and full-grass woodlands. Again, there is no "right answer" to a clustering problem, only substantive and contextual assessment can help establish that a particular grouping is the most satisfactory for a given purpose.

Areas of more consistent response to management.

In our third analysis, we have again used CART to do a supervised clustering analysis. However, this time the goal was not to simplify or generalize a given classification. Here the task is clearly to discover a relationship between a classification and a group of extraneous attributes, i.e. attributes that were not involved in the derivation of that classification. Using CART to predict physiognomic groups from abiotic variables, we are choosing a particular set of plot clusters: those that maximize the relationship between physiognomic groups and abiotic factors. Another difference with the second analysis is that in this case we are also choosing among the abiotic factors, retaining only those that have a significant relationship with the physiognomic groups. This task was best accomplished using supervised clustering.

The purpose of this analysis was to delineate zones where management can expect consistent response. Our assumption is that within zones that are relatively homogeneous in the effect of abiotic factors on hardwood vegetation structure (our abiotic domains), the response of the vegetation to disturbance and management would be more consistent, and thus more predictable. We did not expect to find abiotic domains were all hardwood rangelands would have the same

physiognomy. To refer back to Major's factorial equation, this would mean that biota (including humans) and time (for successional changes) have no influence. To the contrary, we expected to find particular subsets of our physiognomic groups, in contrasting combinations, in each of the abiotic domains. These expectations were met by the set of domains we delineated for the region north and east of the Central Valley: a sequence of clusters with clear shifts in mix of physiognomies. This sequence is corroborated by a concurrent sequence of shifts in species constancy. We also found a good agreement between the precipitation zones delineated by our analysis of the VTM plots and the maps of hardwood rangelands established sixty years later from remotely sensed data. Standiford et al. 1991 found a similar shift in species composition along an elevation-precipitation gradient in the south half of the Sierra Nevada. More importantly, these authors also found many more (321.2) *Q. Douglasii* saplings per hectare at the high end of the precipitation gradient (560 to 660 mm/year) than in the zones below (e.g. 39.2 saplings/ha for 430-560 mm/year). This indicates that the level of precipitation that separates our dry and mesic domains (at 600 mm) corresponds to a major shift in regeneration potential, which has important consequences for management. McClaran and Bartolome (1989) placed at 500 mm/year the limit below which 50% canopy cover *Q. Douglasii* has no depressive effect on forage production, another interaction of importance for management. Further work would be needed to determine if some other threshold values used to delineate our abiotic domains correspond to marked changes in plant response or interactions. Potential change in response of interest for management include tree resprouting rates, potential for shrub cover, oak growth rates, etc.

The scale of an ecological study is determined by the size, extent and resolution of the observations in time and space, and ecologists should be conscious and explicit about the scales they use (Hoekstra et al., 1991). Given the resolution of most of our abiotic data, the vegetation domains identified in this analysis, and the S&T models specified for these domains, are valid only at the regional scale. That is, they are appropriate for tasks such as bioregional planning, monitoring and assessment. They are too general for management at smaller scales, in particular at the ranch scale. During the construction of the S&T model for domain A, it appeared difficult to link two of the physiognomic groups ([15] and [52]) to the others in a single model (Fig. 16). Our hypothesis is that these two groups (and possibly some forms of group [2]) constitute of a second, and mostly

independent, system of vegetation dynamics in the same domain. The fact that these groups are a large proportion of the groups found in domain B&D, which is next in the sequence of increasing yearly precipitation, adds credit to this hypothesis. It is likely that factors such as aspect and topographic position locally modify the general conditions defined by our analysis. These local modifications create pockets in domain A where conditions are similar to those of domain B&D. Further analysis, with higher resolution data, could reveal such patterns occurring at the landscape scale.

Our study also emphasizes the need for zonations that are specific to a particular question. Although we tried using three zonal classifications of California in combination with the other abiotic factors (Table 2), their inclusion produced unsatisfactory sets of clusters and they were eliminated from later analyses. The plant climates zonation of California was developed for crops and ornamental plantings that are not dependent on natural precipitation (Kimball and Brooks, 1959). In the region north and east of the Central Valley, three main plant climates are distinguished: "Cold air basins" by the valley floor, "Thermal belts" at the lower parts of the foothills, and "Digger pine belt" at higher altitude in the foothills. Although these zones were derived from temperature regimes (e.g. the thermal belt is safe for citrus plantations), they happen to reflect in part the changes in yearly precipitation due to orographic effects. CART readily selected the plant climate zone because it reflected both some temperature and precipitation information and then selected other variables to refine the decision tree. However, the relationship between leaf nodes and the physiognomic groups (as judged by the misclassification rate) was weaker and the resulting clusters were less satisfactory than those presented here. This shows that CART variable selection has some of the imperfections of stepwise procedures: once a variable has been introduced into the model, predictors are judged for predictive ability achieved in combination with the preceding variable. Other combinations of variables might exist that are better. We found that the misclassification rate was a useful indicator in selecting a model. It reflects the strength of the relationship between the "predicted" classification and the attributes selected for the tree, even if it is not a measure of cluster goodness. The other two zonations (Koeppen climates and ecological units of California) suffered from the same problem: established for a different purpose, they carried enough relevant information to obfuscate the analysis but not enough to serve the purpose at hand.

Building state-and-transition models.

The construction of S&T models was limited to a single domain because of the difficulty of repeatedly mobilizing a group of experts without appropriate institutional resources. However limited, this step provided evidence for the usefulness of our approach. The physiognomic groups constituted a good starting point for the delineation of vegetation states. The workshop participants had to add only two states to specify a system that reflected their understanding of the vegetation dynamics in the area. The first, a treeless state (state I) could not have been derived from a database of plots populated by oak trees. It is probable that other states not dominated by oaks may become necessary in other domains (e.g. a mixed-coniferous forest state in domain F). The second (state II) came from a differentiation between two levels of tree density in group [28], which was necessary to mark a threshold in dynamics. The possibility that a second set of states exists in this domain (as discussed above) is another interesting outcome of the expert workshops. Here, the favored hypothesis is that there are zones in domain A where vegetation states and dynamics are similar to those in domain B&D. However, we cannot exclude at this point that there may be some lesser-known transitions between these states. In this case, and because of grouping in domains, the elaboration of S&T models plays a role in hypothesis generation.

Two fundamental problems.

Two fundamental problems appeared during the expert workshops that limited the potential use of the resulting S&T models for planning and evaluating management scenarios. The first problem was that the experts followed a prescriptive rather than predictive approach. The following difficulties appeared during the formulation of transitions: how to account for uncertainty in causation (is this factor a cause of the transition) and in the consistency of result (how often does this factor results in actual transition). For example, the workshop participants agreed that grazing (or, more precisely, browsing by domestic grazers) was a factor in the transition from a grassy savannah state without any seedlings and saplings to a state having some advance regeneration. However, they could not specify the timing, conditions or level of livestock grazing (if any) that would either allow or block such a transition. Worse, each participant knew of exceptions: cases where complete exclusion of grazers did not result in appearance

of young trees in the understory, and cases where grazing by livestock did not suppress advance regeneration.

Although they were reluctant to make any general *prediction* about the effect of grazing on regeneration, most participants would be willing to *prescribe* grazing practices that are favorable to oak regeneration. They would go along with the *Guidelines for Managing California's Hardwood Rangelands* (Tinnin, 1996): "Some research studies have shown how grazing management can be applied to actually encourage the development of young seedlings. These studies have shown that early season grazing, with cattle removed from the area prior to the drying up of the annual forages, actually improves moisture available to the developing seedlings and results in higher rates of growth. This grazing activity also reduces the habitat available for rodents which may be a major source of seedling depredation. The same grazing studies also show that if cattle are left on an area late into the spring and summer, that they will preferentially seek out the young oaks, which are often the only green plants on the site."

Thus, it appears that a prescriptive approach to summarization of knowledge would yield a different and more comprehensive S&T knowledge base than a strictly predictive approach. Westoby, Walker and Noy Meir (1989) definitely made room for prescription in the formulation of S&T models. They proposed the S&T formulation because it is "a practicable way to organize information for management." For them, each entry in the catalog of transition "would summarize knowledge about the conditions which induce the transition." This is particularly apparent in their inclusion of a catalog of opportunities — circumstances under which management action such as fire, heavy grazing, removal of grazing, etc. can produce a "favorable" transition — and of hazards — climatic circumstances under which failure to burn, heavy grazing, etc., could produce an "unfavorable" transition.

Obviously, the approach taken in building a S&T knowledge base has repercussion on its fitness for particular applications. Some researchers have taken a purely predictive approach to building state and transition models. Typical of this approach, Scanlan (1994) argued that discrete-time and continuous-time Markov models should be used to describe state and transition models given transition probabilities or transition rates matrices. In this scheme, several transition matrices must be used to reflect different sets of conditions such as dry versus wet years. Such transition matrices could be derived from logistic

regression models. Some participants in our workshops proposed to use regeneration probabilities from logistic models developed for particular conditions and locations. However, such an approach is problematic. A major difficulty lays in the generalization of numbers obtained at precise locations in space and time, and for a specific combination of factors. Another complication is that a model based on probabilities is not deterministic, which complicates interpretation of its results.

The second problem was that the experts followed a goal-oriented rather than exhaustive approach to the summarization of knowledge about transitions. Exhaustivity in the consideration of causal relationships reinforces the generality of a model, and helps avoid brittleness. To pursue the example about oak regeneration, the adoption of favorable livestock grazing practices may not result in the emergence of a cohort of young oaks if other factors are involved that are suppressing regeneration. Such factors include a high impact of wild browsers such as deer, a high level of acorn predation, or competition from an unpalatable weedy understory species. The ideal model would be exhaustive in identifying the possible combinations of factors and spelling out their effect on the different transitions. In the case of hardwood rangelands, the model would combine various levels of intensity and/or seasonality of grazing, browsing, fire, tree cutting and climatic events and spell out the effect (or lack thereof) of these combinations on each state. A systematic way to achieve this would be to build matrices of interactions between main factors for each of the states. These matrices would contain the effect of each particular interaction, if any, in terms of the possibility to trigger or to block a transition. An alternative would be to define lists of global circumstances that influence basic processes such as oak resprouting, seedling germination, browsing intensity, tree mortality, tree growth. In turn, the state of these basic processes would serve to define the actual transitions.

Supervised conceptual clustering.

Artificial experiments on simple datasets that have obvious structure and, unlike real-world data, are noise free can help one understand the distinctive abilities of supervised conceptual clustering. For a detailed example see Vayssières (1998).

Such experiments demonstrate that for conceptual clustering to work, there needs to be to be some potential archetypes for new groups in the set of classes

used to direct (that is, supervise) the clustering. These archetypal classes need not be entirely composed of individuals that will themselves end up in the new cluster. We have seen for instance that only 75% of plots with cover type 28 entered physiognomic group [28]. What is necessary is a strong enough relationship between a majority of the components of some of the classes used to supervise the clustering (the *Y*-variable) and the attributes used to delineate the clusters (the ***X***-variables). Indeed, if all the cover types comprised a similar mix of plot physiognomies, our supervised clustering into physiognomic groups would have been unsuccessful. Similarly, the grouping of physiognomic groups into domains required significant relationships between the physiognomic groups and some of the abiotic variables.

To test this need for a relationship, we tried to redo our third analysis after replacing the abiotic variables with random variables having the same range of values: CART failed to select a decision tree after cross-validation. This is not because the set of random attributes completely lacked structure; CART was able to grow a 222 nodes tree that predicted correctly the physiognomic group of 60 % of the 1177 plots. However, during cross-validation this 222 node tree was pruned back to a single node. This ability to reject overfitting or completely spurious trees through cross-validation is one of the distinctive qualities of CART when it is used for supervised classification (Verbyla, 1987). We have demonstrated that it is also a very useful feature when using CART for supervised clustering. Without this safeguard, we could have selected one of the decision trees built from the random attributes and tried to interpret its leaf nodes in terms of new clusters. For both of the supervised clustering analyses reported in this paper, we selected the decision tree with lowest cross-validated error. This prevented us from capitalizing on chance relationships and thus ensured that our results were applicable to new data. Whereas a lack of relationship between the *Y*-variable and the ***X***-variables renders supervised clustering inoperable, a perfect or near perfect relationship makes supervised clustering superfluous. When we tried to replace one of the abiotic variables in our third analysis with a linear combination of the physiognomic groups 'codes, CART produced a decision tree with 26 nodes that had perfect predictive ability, even through cross-validation.

CART was not intended for supervised clustering but for supervised classification. However CART has many characteristics that make it the best current tool for supervised clustering. Again, CART tree pruning procedure avoids

overfitting the data and allows to work with noisy, real-world data. The gini diversity index used for our analyses is probably the most consistent of the impurity criteria available for problems with a medium to large number of classes (Breiman, 1996). As we have seen, it is also directly related to the concept of profile that we used to compare and eventually aggregate leaf-nodes in the decision tree. The concept of profile is at the basis of correspondence analysis (Greenacre, 1993). We have used correspondence analysis plots to examine graphically the relationships between our new groups and the partition used to supervise the clustering. However, more quantitative means of discriminating among the new groups can be found in an extension of correspondence analysis (Greenacre, 1988). This method allows for testing differences amongst any grouping of the rows (or the columns) of a contingency table. It would be interesting to combine such method with CART to help decide if the aggregation of two (or more) leaf nodes into a single cluster is warranted. The result would be a tool truly dedicated to supervised clustering.

Conclusions.

It is important to delineate zones where particular management decisions can be expected to produce similar results. This study is a step in this direction. However, much work remains to be done in particular in the validation of such zones with independent data sets.

The process of grouping vegetation units into physiognomic groups and abiotic domains proved to be a good way to provide a quantitative basis to the construction of state-and-transition models for California's hardwood rangelands. This process reduces the set of potential vegetation states to a manageable size and help workshop participants focus on a set of transitions operating within a more homogenous subset of environmental influences. However, the development of state-and-transition models through expert workshops will require the involvement of an organization capable of mobilizing groups of experts for the necessary amount of time. It is also clear that the models designed through the workshops should be tested by land-managers and end-users, and revised as necessary. The necessity of directing the knowledge acquisition process to achieve a more exhaustive summarization of causal relationships involved in transitions became apparent in our study. Although this is a fundamental problem, procedural solutions can be found. Making workshop participants made aware of it will also be an important step toward a solution. The other fundamental problem we encountered during our workshops appears is much harder to solve. Difficulties in accounting for uncertainty in causation and consistency of results lead experts toward a prescriptive rather than predictive approach. Such an approach limits the potential of state-and-transition models to be implemented into spatially explicit simulation models for planning and evaluating management scenarios. However, state-and-transition models remain the best framework to date to summarize knowledge in a manner that supports a pro-active land management.

Present trends in vegetation science can be summarized under five headings: formalism, pluralism, functionalism, pragmatism and indeterminism (Mucina, 1997). The approach we have taken in this research reflects several of those tendencies. It is a definitely a pluralist approach, exploring data structure from various angles, and recognizing that there are a multiplicity of ecologically sound classifications that reflect the multi-layered variability of vegetation. Supervised clustering is a particularly interesting tool in this context because it allows

constructing classifications centered on one layer of variability under the supervision of a classification derived from another level of variability. Our approach is also a functionalist one. The limitations of using a floristic approach for summarizing knowledge about vegetation dynamics has led us to develop a classification based on physiognomy. Although more sophisticated schemes based on guilds or functional groups are very promising, the information necessary to develop them is just not available. However, we have shown that the physiognomic approach was valuable in helping separate the influence of biotic and abiotic factors on vegetation, and thus in focusing on vegetation response to management. Lastly, our approach is pragmatic, that is, characterized by using those classification variables (including non-floristic ones), that relate directly to a practical objective such as the derivation of nature-management units. Naturally, such an approach will produce results that have only local validity: universality is sacrificed for applicability towards land management and bioregional planning.

References.

Allen, B. H., Evett R.R., Holzman B. A. & Martin A. J. 1989. Report on rangeland cover type descriptions for California hardwood rangelands. California Department of Forestry and Fire Protection, Forest and Rangeland Resources Assessment Program, Sacramento, CA. USA.

Allen, B. H., Holzman B. A. & Evett R. R. 1991. A classification system for California's hardwood rangelands. Hilgardia 59: 1-45.

Allen-Diaz, B. H. & J. W. Bartolome. (in press) Sagebrush-grass vegetation dynamics: comparing classical and State-Transition models. Ecological Applications.

Anderberg, M. R. 1973. Cluster Analysis for applications. Academic Press, New York, USA.

Austin, M. P.,. Nicholls A. O & Margules C. R. 1990. Measurement of the realized qualitative niche: environmental niches of five Eucalyptus species. Ecological Monographs 60:161-177.

Bartolome, J.W., Muick P.C. and M.P. McClaran. 1987. Natural regeneration of Californian hardwoods. In: Plumb, T.R. and N.H. Pillsbury, tech. coords. Proceedings of the symposium on multiple use management of California's hardwood resources; November 12-14, 1986, San Luis Obispo, Ca. Gen. Tech. Rep. PSW-100. Berkeley, Ca: Pacific Southwest Forest and Range Experiment Station: Forest Service, U.S.D.A.: 26-31.

Bellamy, J. A. & Brown J. R. 1994. State and transition models for rangelands .7. Building a state and transition model for management and research on rangelands. Tropical Grasslands 28:247-255.

Breiman, L., Friedman J. H., Olshen R.A. & Stone C. J. 1984. Classification and regression trees. The Wadsworth statistics /probability series, Chapman and Hall, Inc. New York, New York, USA.

Breiman, L. 1996. Some properties of splitting criteria. Machine Learning, 24:41-47.

Briscoe, G. & Caelli T. 1996. A compendium of machine learning. Vol. 1 Symbolic machine learning. Ablex Publishing Corporation, Norwood, NJ, USA.

Buntine, W. & Niblett T. 1992. A further comparison of splitting rules for decision-tree induction. Machine Learning, 8:75-85.

Bork, E. W., Hudson R. J. & A. W. Bailey. 1997. Upland plant community classification in Elk Island national park, Alberta, Canada, using disturbance history and physical site factors. Plant Ecology, 130:171-190.

Clements, F.E. 1916. Plant succession: an analysis of the development of vegetation. Carnegie Inst. Washington Pub. 242: 1-512.

Connell, J.H. and R.O. Slatyer. 1977. Mechanisms of succession in natural communities and their role in community stability and organization. American Naturalist 111: 1119-1144.

Duffie, J. A., & Beckman W. A. 1980. Solar engineering of thermal processes. Wiley, New York, New York, USA.

Dykersthuis, E.J. 1949. Condition and Management of Rangeland based on quantitative ecology. Journal of Range Management 2: 104-115.
Dykersthuis, E.J. 1958. Ecological principles in range evaluation. Botanical Review 24: 253-272.
Evett, R. R. 1994. Determining environmental realized niches for six oak species in California through direct gradient analysis and ecological response surface modeling. Ph.D. Dissertation. University of California Berkeley, California, USA.
Fernández Alés, R., Laffarga J.M. and F. Ortega. 1993. Strategies in Mediterranean grassland annuals in relation to stress and disturbance. Journal of Vegetation Science, 4: 313-322.
Fisher, D. H. 1987. Knowledge acquisition via incremental conceptual clustering. Machine Learning. 2:139-172.
Friedel, M.H. 1991. Range condition assessment and the concept of thresholds: a viewpoint. Journal of range management, 44: 422-426.
Friedel, M.H., Bastin G.N. and G.F. Griffin. 1988. Range assessment and monitoring in arid lands: the derivation of functional groups to simplify vegetation data. Journal of Environmental Management, 27: 85-97.
George, M.R., Brown J.R. and W.J. Clawson. 1992. Application of non-equilibrium ecology to management of Mediterranean grasslands. J. Range Manage. 45: 436-440.
George M.R., Vayssières, M.P., Bernheim L., Young j. & R.E. Plant. 1993. An Intelligent GIS for Rangeland Impact Assessment. A report prepared for: Strategic Resources and Planning Program. California Department of Forestry and Fire Protection, Sacramento, CA. USA.
Goudey C. B. & Smith D. W. Ecological units of California : subsections. 1994. USDA Forest Service, Pacific Southwest Region Geometronics. United States Department of Agriculture, Forest Service, Pacific Southwest Region, in cooperation with Natural Resource Conservation Service. Washington DC, USA.
Greenacre, M. J. 1988. Clustering the rows and columns of a contingency table. Journal of Classification, 5:39-51.
Greenacre, M. J. 1993. Correspondence analysis in practice. Academic Press Inc. San Diego, CA. USA.
Griffin, J.R. & Critchfield W. B. 1972. The distribution of forest trees in California. Research paper PSW-82. USDA Forest Service, Pacific Southwest Forest and Range Experiment Station, Berkeley, California, USA.
Godron, M. and R.T.T. Forman. 1983. Landscape modification and changing ecological characteristics. In: Mooney H.A. and M. Godron (eds.). Disturbance and Ecosystems. Ecological Studies no. 44. Springer-Verlag, Berlin.
Griffin, J.R., McDonald P.M. and P.C. Muick. 1987. California oaks : a bibliography. Pacific Southwest Forest and Range Experiment Station, Berkeley, California.
Hand, D. J. 1997. Construction and Assessment of Classification Rules. John Wiley & Sons Ltd, Chichester, U.K.

Hill, M. O. 1979a. TWINSPAN –A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of individuals and attributes. Cornell University, Ithaca, NY, USA. 90 pp.
Hill, M. O. 1979b. DECORANA –A FORTRAN program for detrended correspondence analysis and reciprocal averaging. Cornell University, Ithaca, NY, USA. 52 pp.
Hoekstra, T. W., Allen, T. F. H. & Flather, C. H. 1991. Implicit scaling in ecological research. Bioscience, 41: 148-154.
Holling, C.S. 1973. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics, 4: 1-23.
Holzman, B.A. and B.H. Allen-Diaz. Vegetation change in blue oak woodlands in California. In: Standiford R.B., tech. coord. Proceedings of the symposium on oak woodland and hardwood rangeland management; October 31 - November 2, 1990; Davis, Ca. Gen. Tech. Rep. PSW-126. Berkeley, Ca: Pacific Southwest Forest and Range Experiment Station: Forest Service, U.S.D.A. 1991: 189-193.
James, J. W. 1966. A modified Koeppen classification of California ŝ climates according to recent data. The California Geographer, 7:1-12.
Jameson, D.A. 1991. Effects of single season and rotation harvesting on cool-and-warm-season grasses of a mountain grassland. Journal of Range Management, 44: 327-329.
Jensen, H.A. 1947. A system for classifying vegetation in California. California Fish and Game 33:199-266.
Jenny, H. 1941. Factors of soil formation. McGraw-Hill, New York.
Jenny, H. 1961. Derivation of state factor equations of soils and ecosystems. Proceedings of the Soil Science Society of America, 25: 385-388.
Jones, R.M. 1992. Resting from grazing to reverse changes in sown pasture composition: application of the ŝtate-and-transition ’model. Tropical Grasslands, 26: 97-99.
Kass, G. V. 1980. An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29:119-127.
Keddy, P.A. 1989. Competition. Chapman and Hall, London.
Keddy, P.A. 1992. Assembly and response rules: two goals for predictive community ecology. Journal of Vegetation Science, 3: 157-164.
Kimball, M. H. and F. A. Brooks. 1959. Plantclimates of California. California Agriculture, 13(5):7-12.
Laycock, W.A. 1989. Secondary succession and range condition criteria: introduction to the problem. pp. 1-15. In: Lauenroth, W.K. and W.A. Laycock (eds.) Secondary succession and the evaluation of rangeland condition. Westview Press, Boulder, Colorado.
Leishman, M.R. and M. Westoby. 1992. Classifying plants into groups on the basis of associations of individual traits --evidence from Australian semi-arid woodland. Journal of Ecology, 80: 417-424.
Maclaran , M. P. & Bartolome J.W. 1989. Effects of *Quercus Douglasii* (Fagaceae) on herbaceous understory along a rainfall gradient. Madroño, 36:141-153.
Macnaughton-Smith, P. 1963. The classification of individuals by the possession of attributes associated with a criterion. Biometrics, 19:363-366.

Macnaughton-Smith, P. 1965. Some statistical and other numerical techniques for classifying individuals. Home Office Research Unit Report No. 6. Her Majesty's Stationery Office, London, United Kingdom.
Major, J. 1951. A functional, factorial approach to plant ecology. Ecology, 32(3): 392-411.
Major, J. and W.T. Pyott. 1966. Buried viable seeds in two California bunchgrass sites and their bearing on the definition of a flora. Vegetatio, 13: 253-282.
Matthews, G. & Hearne J. 1991. Clustering without a metric. IEEE Transactions on Pattern Analysis and Machine Intelligence 13(2):175-184.
Matthews, G., Matthews R. & Landis W. 1995. Nonmetric conceptual clustering in ecology and ecotoxicology. AI applications. 9:41-48.
May, R.M. 1977. thresholds and breakpoints in ecosystems with a multiplicity of stable states. Nature, 269: 471-477.
Mayer, K. E. & Laudenslayer W. F. Jr. (eds.). 1988. A guide to wildlife habitats of California. California Department of Forestry and Fire Protection, Sacramento, CA. USA.
Michalski, R. S. & Stepp R. 1982. Revealing conceptual structure in data by inductive inference. In: J.E. Hayes, Michie D. & Pao Y. H. (eds.) Machine Intelligence 10. Halsted Press, John Wiley & Sons, New York, NY, USA.
Michalski, R. S. & Stepp R. 1983. Learning from observation: Conceptual clustering. In: R. S. Michalski, Carbonell J., & Mitchell T. M. (eds.) Machine Learning: An Artificial Intelligence Approach. Tioga, Palo Alto, CA., USA.
Mingers, J. 1989. An empirical comparison of selection measures for decision-tree induction. Machine Learning, 3:319-342.
Milligan, G. W. 1996. Clustering validation: results and implications for applied analyses. In: P. Arabie, L. J. Hubert & G. De Soete (eds.), Clustering and Classification, World Scientific Publishing Co., River Edge, NJ, USA.
Mirkin, B. 1996. Mathematical Classification and Clustering. Kluwer Academic Publishers, Dordrecht, The Netherlands.
Morgan, J. N. &Messenger R. C. 1973. THAID, a sequential analysis program for the analysis of nominal scale dependent variables, by James N. Morgan [and] Robert C. Messenger. Ann Arbor, Survey Research Center, Institute for Social Research, University of Michigan. USA.
Mucina, L. 1997. Classification of vegetation: Past, present and future. Journal of Vegetation Science, 8:751-760.
Munz, P.A. & Keck D. D. 1968. A California Flora with supplement. University of California Press, Berkeley and Los Angeles, CA, USA.
Noble I.R. and R.O. Slatyer. 1980. The use of vital attributes to predict successional changes in plant communities subject to recurrent disturbance. Vegetatio, 43: 5-21.
Noy-Meir, I. 1982. Stability of plant herbivore models and possible applications to savanna. In: Huntley B.J. and B.H. Walker (eds.) Ecology of tropical savannas. Springer-Verlag, Berlin. pp. 591-609.
Pacific Meridian Resources. 1994. California Hardwood Rangeland Monitoring Final Report. California Department of Forestry and Fire Protection, Forest and Rangeland Resources Assessment Program, Sacramento, CA. USA.

Pavlick, B.M., Muick P.C., Johnson S.G. and M. Popper. 1991. Oaks of California. Cachuma Press, Los Olivos, CA.

Pillsbury, N., De Lasaux M., Pryor R. & W. Bremer. 1991. Mapping and GIS database development for California's hardwood resources. California Department of Forestry and Fire Protection, Forest and Rangeland Resources Assessment Program, Sacramento, CA. USA.

Plant, R. E., Vayssières M. P., Greco S. E., George M. R. & T. E. Adams. [In Press]. A qualitative spatial model of hardwood rangeland State-and-transition vegetation dynamics. Journal of Rangeland Management.

Plumb, T.R. 1980. Response of oaks to fire. In: Plumb, T.R., tech. coord. Proceedings of the symposium on the ecology, management and utilization of California Oaks; 1979 June 26-28, Claremont, Ca. Gen. Tech. Rep. PSW-44. Berkeley, Ca: Pacific Southwest Forest and Range Experiment Station: Forest Service, U.S.D.A.: 202-215.

Powell, R. Electronic data processing codes for California wildland plants. 2nd edition. University of California, Davis, CA, USA.

Quinlan, J. R. 1986. Induction of decision trees. Machine Learning, 1:81-106.

Quinlan, J. R. 1993. C4.5 : Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California, USA.

Robinson, A. H., Sale R. & Morrison J. 1978. Elements of cartography (4th ed.). John Wiley and Sons, New York, NY, USA.

Rostovtsev, P. S. & Mirkin B. G. 1985. Methods for relative hierarchical grouping. In: B. G. Mirkin. Grouping in socio-economical studies: Methods for constructing and analyzing. Finansy i Statistika, Moscow, Russia (in Russian).

Sampson, A.W. 1917. Succession as a factor in range Management. J. Forest. 15: 593-596.

Scanlan, J.C. 1994. State and transition models for rangelands. 5. The use of state and transition models for predicting vegetation change in rangelands. *Tropical Grasslands,* 28: 229-240.

Smith, E.L. 1989. Range condition and secondary succession: a critique. pp. 103-141. In: Lauenroth, W.K. and W.A. Laycock (eds.) Secondary succession and the evaluation of rangeland condition. Westview Press, Boulder, Colorado.

Sonquist, J. A. &Morgan J. N. 1964. The detection of interaction effects; a report on a computer program for the selection of optimal combinations of explanatory variables. Ann Arbor, Survey Research Center monograph no. 35, Institute for Social Research, University of Michigan. USA.

Sonquist, J.A., Baker E. L. & Morgan J. N. 1973. Searching for structure. Institute for social research, Ann Arbor, Michigan, USA.

Stafford Smith, M. and G. Pickup. 1993. Out of Africa, Looking in: Understanding vegetation change. In: Behnke R.H., Scoones I. and C. Kerven (eds.) Range Ecology at Disequilibrium. Overseas Development Institute, London.

Standiford, R., McDougald N., Phillips R. and A. Nelson. 1991. South Sierra oak regeneration weak in sapling stage. California Agriculture, 45: 12-14.

Steinberg, D. and P. Colla. 1995. CART: Tree-structured nonparametric data analysis. Salford Systems, San Diego, CA, USA.

Taush, R.J., Wigand P.E., and J.W. Burkhardt. 1993. Viewpoint: Plant community thresholds, multiple steady states, and multiple successional pathways: legacy of the quaternary. J. Range Manage. 46: 439-447.

Tinnin P. (ed.) 1996. Guidelines for managing California ŝ hardwood rangelands. University of California Division of Agriculture and Natural resources Publication 3368. Oakland, CA. USA.

van Tongeren, O. F. R. 1995. Cluster Analysis. In Jongman, R. H. G., C. J. F. ter Braak, & O. F. R. van Tongeren, editors. Data Analysis in Community and Landscape ecology. New edition. Cambridge University Press, Cambridge, U.K.

Vayssières, M. P. 1998. Physiognomy and spatial structure of California's hrdwood rangelands: recursive partitioning analyses as a basis for state-and-transition models. Ph.D. dissertation, University of California Davis, CA. USA.

Vayssières, M.P., George M.R., Bernheim L., Young j. and R.E. Plant. 1993. An Intelligent GIS for Rangeland Impact Assessment. *Proceedings of the 4th Annual Conference on Artificial Intelligence, Simulation, and Planning in High Autonomy Systems.* Tucson, Arizona. September 20-22, 1993. IEEE Computer Society Press.

Verbyla, D. L. 1987. Classification trees: a new discrimination tool. Canadian Journal of Forest Research 17:1150-1152.

Watt A.S. 1947. Pattern and process in the plant community. J. Ecol. 35: 1-22.

Weaver, J.E. and F.E. Clements. 1938. Plant Ecology. McGraw-Hill, New York.

Westoby, M., Walker, B. and I. Noy-Meir. 1989. Opportunistic management for rangelands not at equilibrium. Journal of Range Management, 42(4): 266-274.

Whittaker, R.H. 1953. A consideration of climax theory: the climax as a population and pattern. Ecological monographs, 23:41-78.

Wieslander, A. E. 1935. A vegetation type map of California. Madroño 3:140-144.

Williams, W. T. 1976. Types of classification. In W.T. Williams ed. Pattern Analysis in Agricultural Science. CSIRO, Melbourne, Australia; and Elsevier Scientific Publishing Co., Amsterdam, The Netherlands.

Williams, W.T. &Lambert J. M. 1959. Multivariate methods in plant ecology. I. Association analysis in plant communities. Journal of Ecology 47:83-101.

Williams, W.T. &Lambert J. M. 1960. Multivariate methods in plant ecology. II. The use of an electronic digital computer for association analysis. Journal of Ecology 48:689-710.

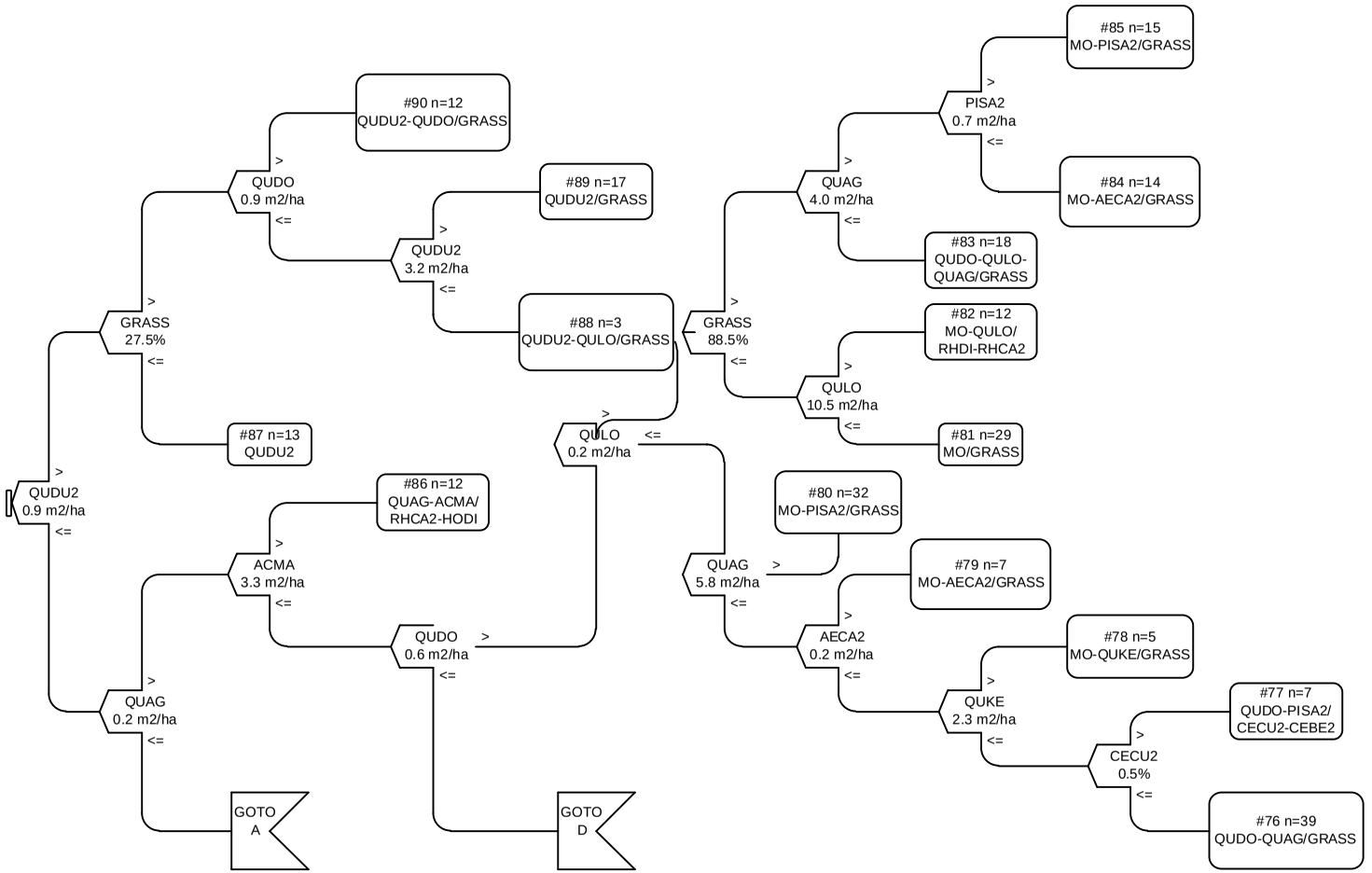
Wilson, A.D. 1989. The development of systems of assessing the condition of rangeland in Australia. pp. 77-102. In: Lauenroth, W.K. and W.A. Laycock (eds.) Secondary succession and the evaluation of rangeland condition. Westview Press, Boulder, Colorado.

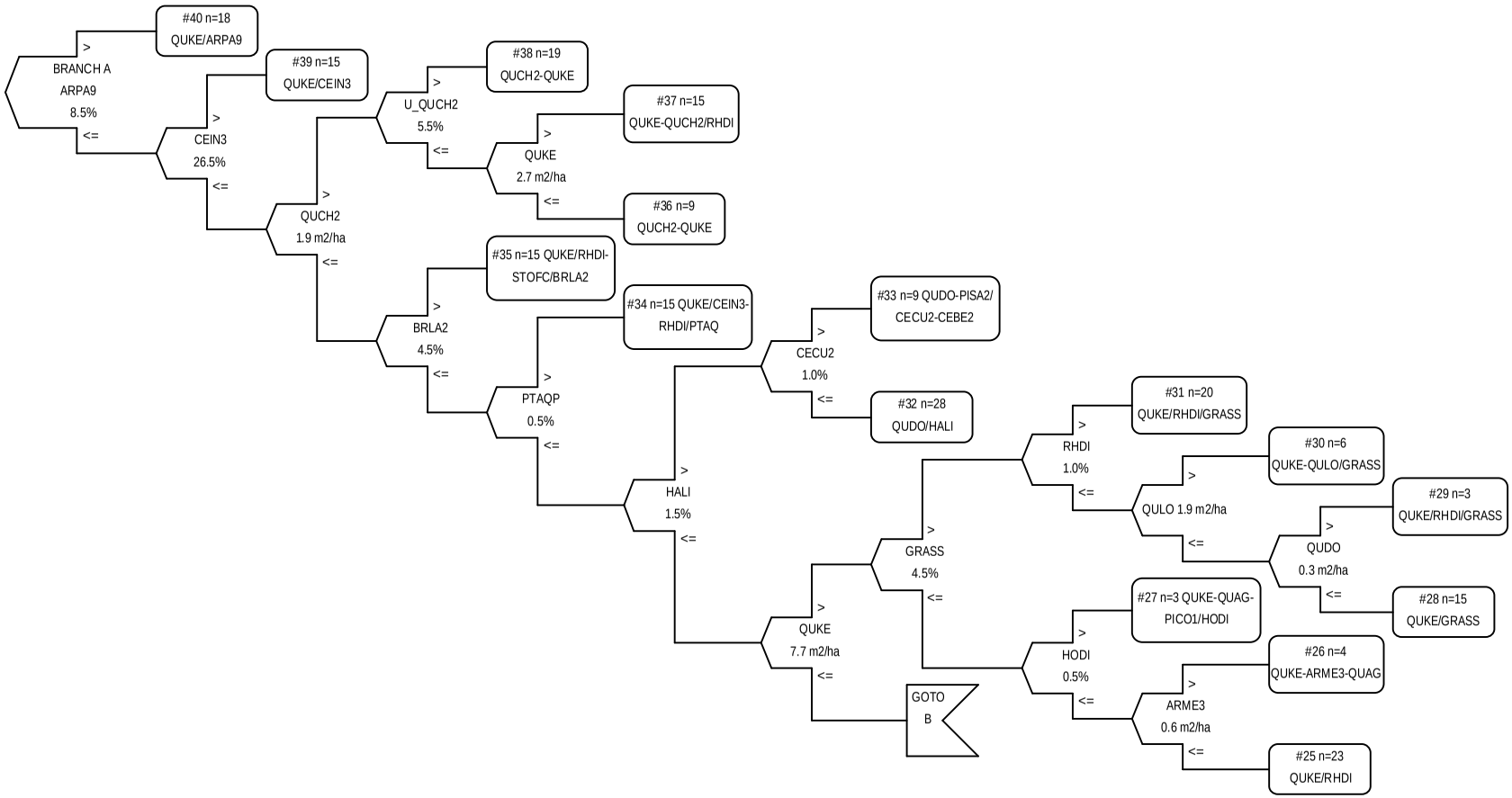
Appendices

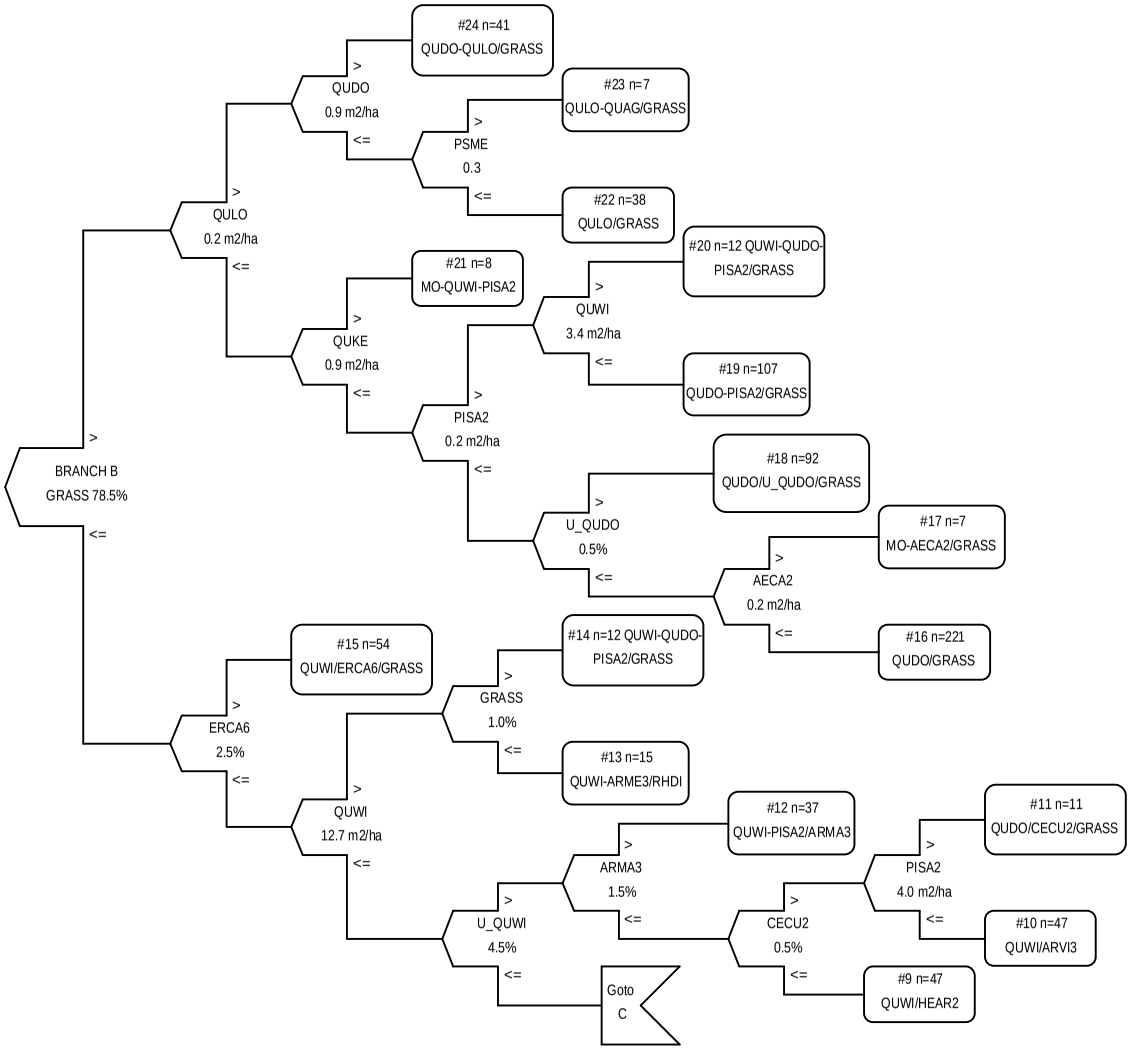
Appendix 1: Plant species codes (Powell, 1987) with scientific names and common names from Munz and Keck (1968).

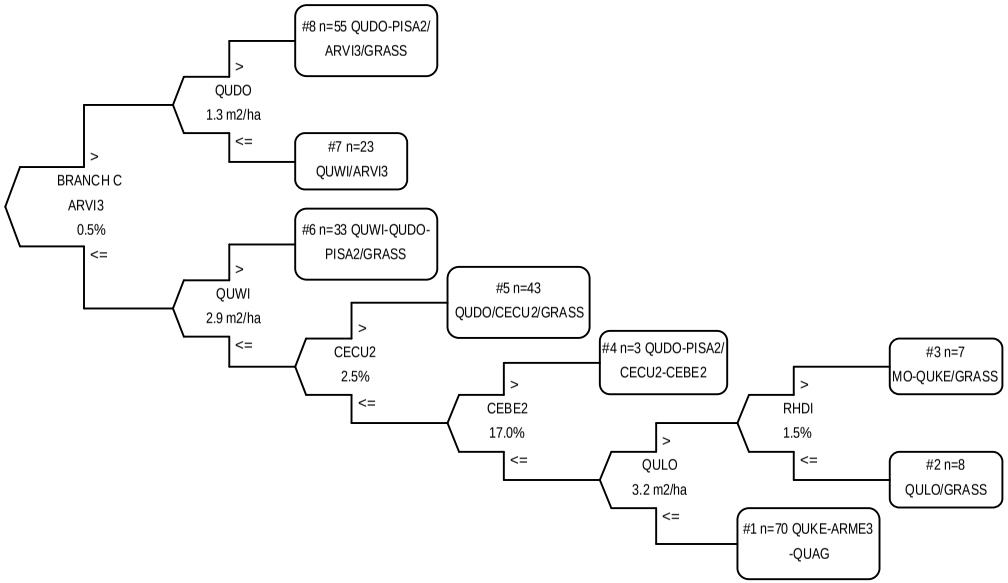
Code	*Genus*	*Species*	*Variety*	*Common Name*
ACMA	*ACER*	*MACROPHYLLUM*	-	Bigleaf maple
ADFA	*ADENOSTOMA*	*FASCICULATUM*	-	Chamise
AECA2	*AESCULUS*	*CALIFORNICA*	-	California buckeye
ARCA7	*ARTEMISIA*	*CALIFORNICA*	-	Coast sagebrush
ARMA3	*ARCTOSTAPHYLOS*	*MANZANITA*	-	Common manzanita
ARME3	*ARBUTUS*	*MENZIESII*	-	Madrone
ARPA9	*ARCTOSTAPHYLOS*	*PATULA*	-	Greenleaf manzanita
ARVI3	*ARCTOSTAPHYLOS*	*VISCIDA*	-	Whiteleaf manzanita
BAPI	*BACCHARIS*	*PILULARIS*	-	Baccharis
BRLA2	*BRODIAEA*	*LAXA*	-	Grass-nut
CEBE2	*CERCOCARPUS*	*BETULOIDES*	-	Birchleaf mountain-
CECU2	*CEANOTHUS*	*CUNEATUS*	-	Wedgeleaf ceanothus
CEIN3	*CEANOTHUS*	*INTEGERRIMUS*	-	Deerbrush
COCO5	*CORYLUS*	*CORNUTA*	-	Hazelnut
ERCA6	*ERIODICTYON*	*CALIFORNICUM*	-	California yerba santa
HALI	*HAPLOPAPPUS*	*LINEARIFOLIUS*	-	Narrowleaf
HEAR2	*HETEROMELES*	*ARBUTIFOLIA*	-	Toyon, christmas
HODI	*HOLODISCUS*	*DISCOLOR*	-	Ocean spray
PICO1	*PINUS*	*CONTORTA*	-	Beach pine
PISA2	*PINUS*	*SABINIANA*	-	Foothill pine
POFR3	*POPULUS*	*FREMONTII*	-	Fremont cottonwood
PSME	*PSEUDOTSUGA*	*MENZIESII*	-	Douglas-fir
PTAQ	*PTERIDIUM*	*AQUILINUM*	-	Bracken fern
QUAG	*QUERCUS*	*AGRIFOLIA*	-	Coast live oak
QUCH2	*QUERCUS*	*CHRYSOLEPIS*	-	Canyon live oak
QUDO	*QUERCUS*	*DOUGLASII*	-	Blue oak
QUDU2	*QUERCUS*	*DUMOSA*	-	California scrub oak
QUKE	*QUERCUS*	*KELLOGGII*	-	California black oak
QULO	*QUERCUS*	*LOBATA*	-	Valley oak
QUWI	*QUERCUS*	*WISLIZENII*	-	Interior live oak
RHCA2	*RHAMNUS*	*CALIFORNICA*	-	California coffeeberry
RHDI	*RHUS*	*DIVERSILOBA*	-	Poison-oak
RUVI2	*RUBUS*	*VITIFOLIUS*	-	Coast california
SAME4	*SALVIA*	*MELLIFERA*	-	Black sage
STOFC	*STYRAX*	*OFFICINALIS*	californica	California storax
SYRI	*SYMPHORICARPOS*	*RIVULARIS*	-	Upright snowberry
UMCA1	*UMBELLULARIA*	*CALIFORNICA*		California bay
GRASS				All grass species
LITTER				Ground litter
MO				Mixed (several) oaks

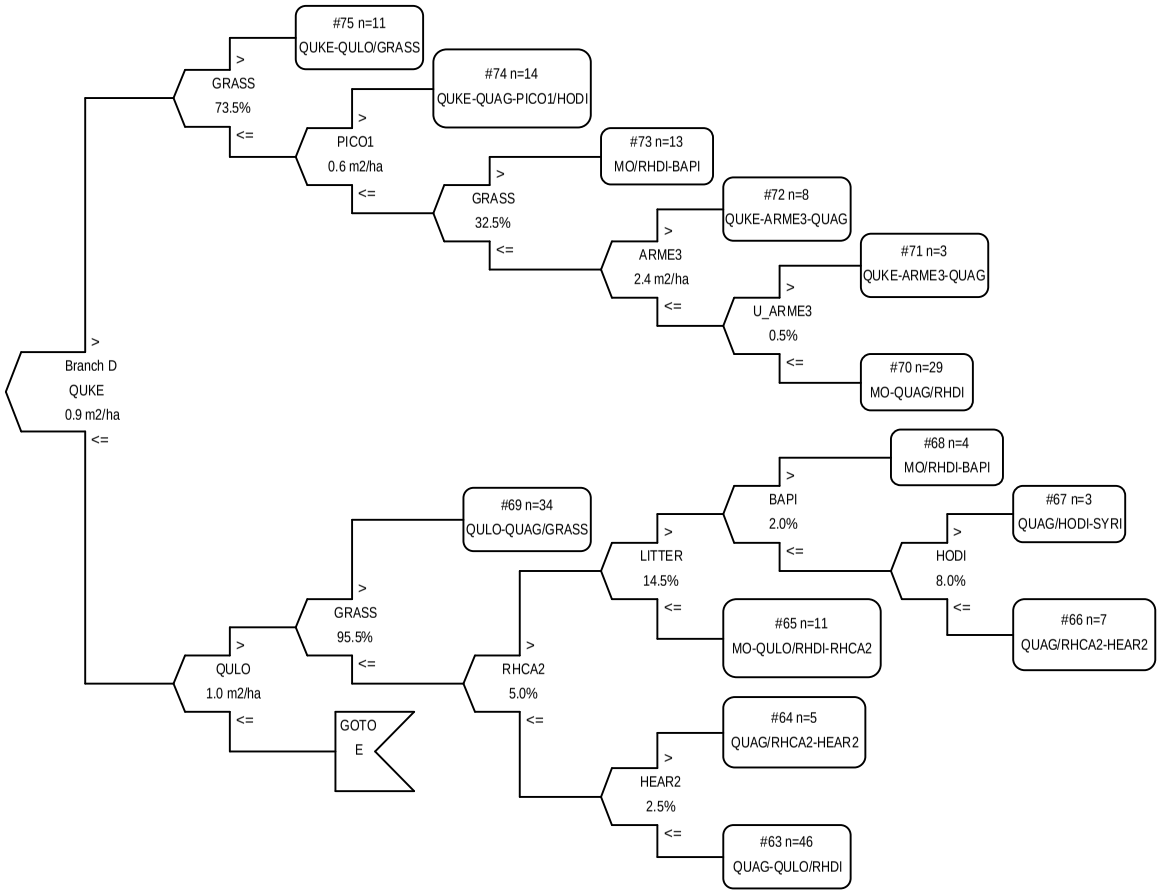
Appendix 2: CART decision tree constituting a key to the cover types of Allen et al. (1991). Note: the tree has been divided into a root branch (this page) and six other branches (following pages).

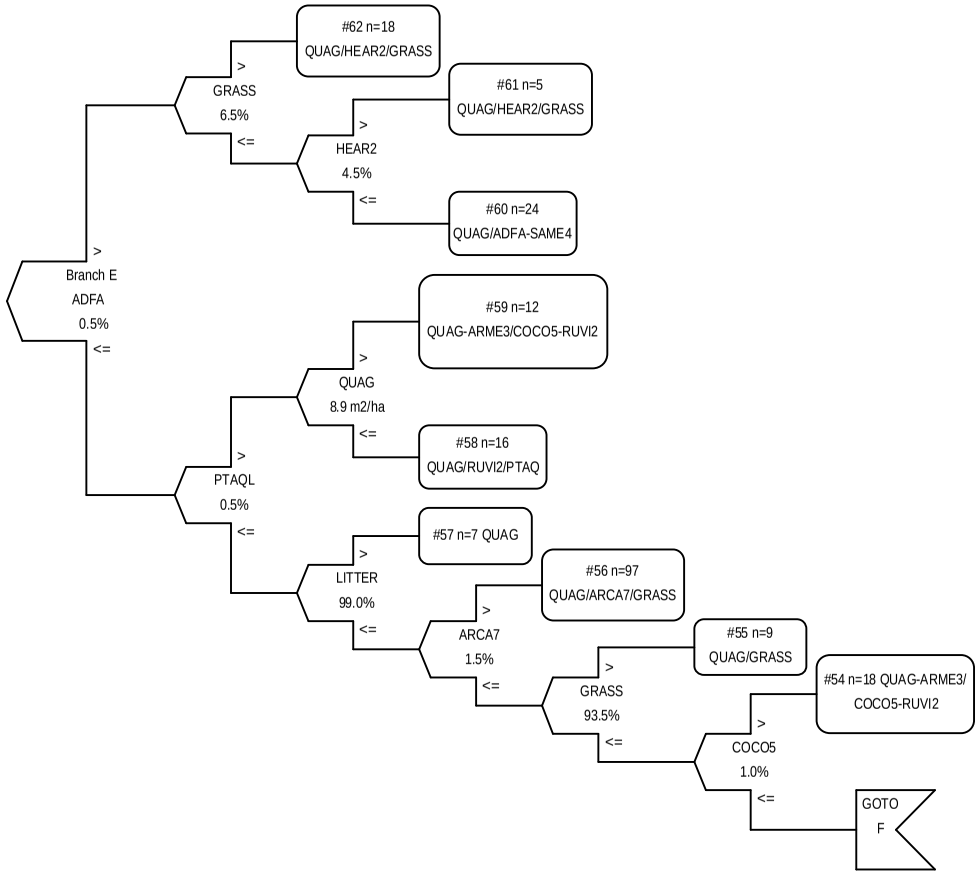


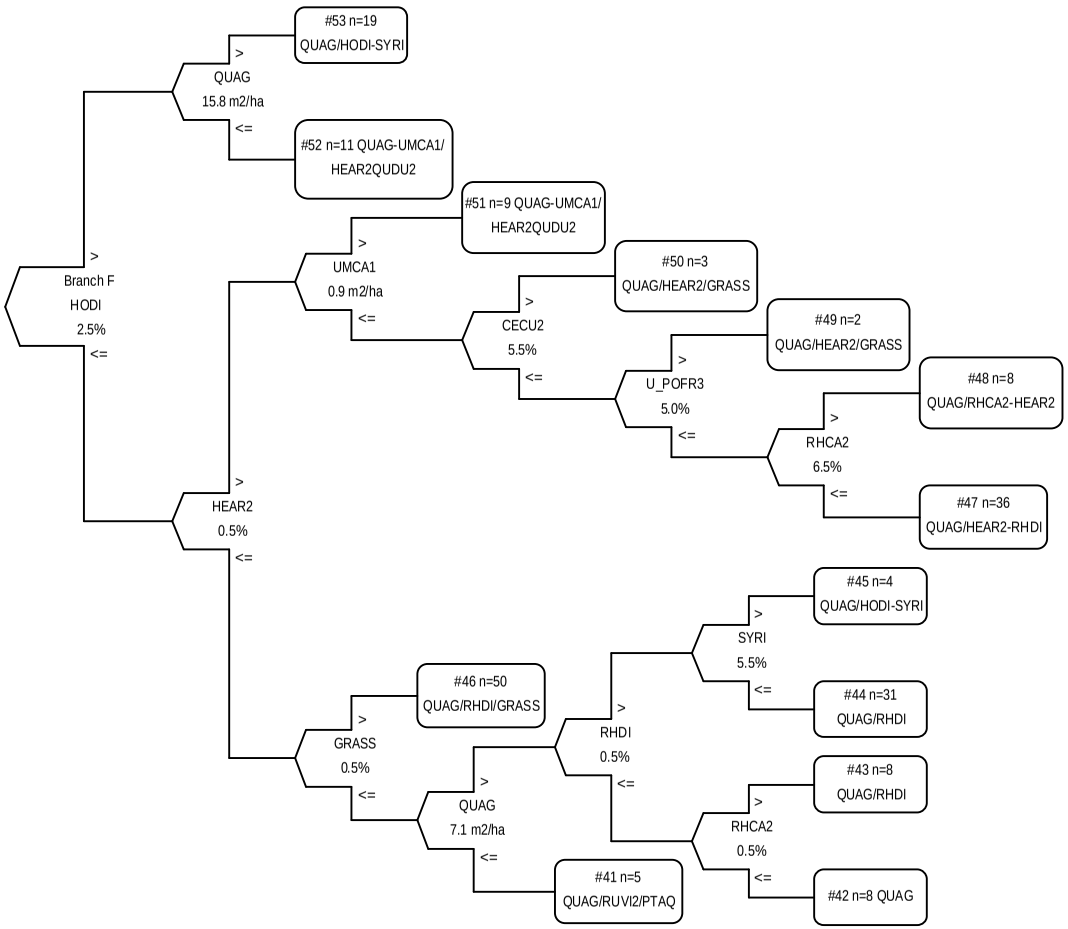

#40 n=18
QUKE/ARPA9
>
BRANCH A
ARPA9
8.5%
<=
#39 n=15
QUKE/CEIN3
>
CEIN3
26.5%
<=
#38 n=19
QUCH2-QUKE
>
U_QUCH2
5.5%
<=
#37 n=15
QUKE-QUCH2/RHDI
>
QUKE
2.7 m2/ha
<=
#36 n=9
QUCH2-QUKE
>
QUCH2
1.9 m2/ha
<=
#35 n=15 QUKE/RHDI-
STOFC/BRLA2
>
BRLA2
4.5%
<=
#34 n=15 QUKE/CEIN3-
RHDI/PTAQ
>
PTAQP
0.5%
<=
#33 n=9 QUDO-PISA2/
CECU2-CEBE2
>
CECU2
1.0%
<=
#32 n=28
QUDO/HALI
>
HALI
1.5%
<=
#31 n=20
QUKE/RHDI/GRASS
>
RHDI
1.0%
<=
#30 n=6
QUKE-QULO/GRASS
>
QULO 1.9 m2/ha
<=
#29 n=3
QUKE/RHDI/GRASS
>
QUDO
0.3 m2/ha
<=
#28 n=15
QUKE/GRASS
>
GRASS
4.5%
<=
#27 n=3 QUKE-QUAG-
PICO1/HODI
>
HODI
0.5%
<=
#26 n=4
QUKE-ARME3-QUAG
>
ARME3
0.6 m2/ha
<=
#25 n=23
QUKE/RHDI
>
QUKE
7.7 m2/ha
<=
GOTO
B


BRANCH B
GRASS 78.5%
>
<=
QULO
0.2 m2/ha
>
<=
QUDO
0.9 m2/ha
>
<=
#24 n=41
QUDO-QULO/GRASS
PSME
0.3
>
<=
#23 n=7
QULO-QUAG/GRASS
#22 n=38
QULO/GRASS
QUKE
0.9 m2/ha
>
<=
#21 n=8
MO-QUWI-PISA2
PISA2
0.2 m2/ha
>
<=
QUWI
3.4 m2/ha
>
<=
#20 n=12 QUWI-QUDO-PISA2/GRASS
#19 n=107
QUDO-PISA2/GRASS
U_QUDO
0.5%
>
<=
#18 n=92
QUDO/U_QUDO/GRASS
AECA2
0.2 m2/ha
>
<=
#17 n=7
MO-AECA2/GRASS
#16 n=221
QUDO/GRASS
ERCA6
2.5%
>
<=
#15 n=54
QUWI/ERCA6/GRASS
QUWI
12.7 m2/ha
>
<=
GRASS
1.0%
>
<=
#14 n=12 QUWI-QUDO-PISA2/GRASS
#13 n=15
QUWI-ARME3/RHDI
U_QUWI
4.5%
>
<=
ARMA3
1.5%
>
<=
#12 n=37
QUWI-PISA2/ARMA3
Goto
C
CECU2
0.5%
>
<=
PISA2
4.0 m2/ha
>
<=
#11 n=11
QUDO/CECU2/GRASS
#10 n=47
QUWI/ARVI3
#9 n=47
QUWI/HEAR2


#8 n=55 QUDO-PISA2/ARVI3/GRASS
>
QUDO
1.3 m2/ha
<=
#7 n=23 QUWI/ARVI3
>
BRANCH C
ARVI3
0.5%
<=
#6 n=33 QUWI-QUDO-PISA2/GRASS
>
QUWI
2.9 m2/ha
<=
#5 n=43 QUDO/CECU2/GRASS
>
CECU2
2.5%
<=
#4 n=3 QUDO-PISA2/CECU2-CEBE2
>
CEBE2
17.0%
<=
#3 n=7 MO-QUKE/GRASS
>
RHDI
1.5%
<=
#2 n=8 QULO/GRASS
>
QULO
3.2 m2/ha
<=
#1 n=70 QUKE-ARME3-QUAG


Branch D
QUKE
0.9 m2/ha
>
<=
GRASS
73.5%
>
<=
#75 n=11
QUKE-QULO/GRASS
PICO1
0.6 m2/ha
>
<=
#74 n=14
QUKE-QUAG-PICO1/HODI
GRASS
32.5%
>
<=
#73 n=13
MO/RHDI-BAPI
ARME3
2.4 m2/ha
>
<=
#72 n=8
QUKE-ARME3-QUAG
U_ARME3
0.5%
>
<=
#71 n=3
QUKE-ARME3-QUAG
#70 n=29
MO-QUAG/RHDI
QULO
1.0 m2/ha
>
<=
GRASS
95.5%
>
<=
#69 n=34
QULO-QUAG/GRASS
GOTO
E
RHCA2
5.0%
>
<=
LITTER
14.5%
>
<=
BAPI
2.0%
>
<=
#68 n=4
MO/RHDI-BAPI
HODI
8.0%
>
<=
#67 n=3
QUAG/HODI-SYRI
#66 n=7
QUAG/RHCA2-HEAR2
#65 n=11
MO-QULO/RHDI-RHCA2
HEAR2
2.5%
>
<=
#64 n=5
QUAG/RHCA2-HEAR2
#63 n=46
QUAG-QULO/RHDI


#62 n=18
QUAG/HEAR2/GRASS
>
GRASS
6.5%
<=
#61 n=5
QUAG/HEAR2/GRASS
>
HEAR2
4.5%
<=
#60 n=24
QUAG/ADFA-SAME4
>
Branch E
ADFA
0.5%
<=
#59 n=12
QUAG-ARME3/COCO5-RUVI2
>
QUAG
8.9 m2/ha
<=
#58 n=16
QUAG/RUVI2/PTAQ
>
PTAQL
0.5%
<=
#57 n=7 QUAG
>
LITTER
99.0%
<=
#56 n=97
QUAG/ARCA7/GRASS
>
ARCA7
1.5%
<=
#55 n=9
QUAG/GRASS
>
GRASS
93.5%
<=
#54 n=18 QUAG-ARME3/
COCO5-RUVI2
>
COCO5
1.0%
<=
GOTO
F


#53 n=19
QUAG/HODI-SYRI
>
QUAG
15.8 m2/ha
<=
#52 n=11 QUAG-UMCA1/
HEAR2QUDU2
#51 n=9 QUAG-UMCA1/
HEAR2QUDU2
>
UMCA1
0.9 m2/ha
<=
#50 n=3
QUAG/HEAR2/GRASS
>
CECU2
5.5%
<=
#49 n=2
QUAG/HEAR2/GRASS
>
U_POFR3
5.0%
<=
#48 n=8
QUAG/RHCA2-HEAR2
>
RHCA2
6.5%
<=
#47 n=36
QUAG/HEAR2-RHDI
>
Branch F
HODI
2.5%
<=
>
HEAR2
0.5%
<=
#45 n=4
QUAG/HODI-SYRI
>
SYRI
5.5%
<=
#44 n=31
QUAG/RHDI
#46 n=50
QUAG/RHDI/GRASS
>
GRASS
0.5%
<=
>
RHDI
0.5%
<=
#43 n=8
QUAG/RHDI
>
QUAG
7.1 m2/ha
<=
>
RHCA2
0.5%
<=
#41 n=5
QUAG/RUVI2/PTAQ
#42 n=8 QUAG

Appendix 3: Catalogs of states and transitions.

Vegetation type: Blue oak (*Quercus douglasii*) dominated savanna and woodland. Blue oak dominant, sometimes associated with foothill pine (*Pinus Sabiniana*), interior live oak (*Q. Wislizenii*) and/or valley oak (*Q. lobata*).
Location: lower part of the Sierra Nevada foothills and adjacent valley floor, California. This zone receives less than 600 mm of yearly precipitation on average (abiotic domain A).

Catalog of States

State I. (group {0}) Annual grassland. Very few if any shrubs or trees.
State II. (group {1}) Savanna, 10-20% canopy closure. Understory composed almost exclusively of annual grasses.
State III. (group [28]) Savanna, 20-60% canopy closure. Understory composed almost exclusively of annual grasses. This is the most common hardwood vegetation state in this zone.
State IV. (group [32]) Savanna , 20-60% canopy. Young *Quercus* trees are present in the understory otherwise composed almost exclusively of annual grasses, with few shrubs (*Ceanothus cuneatus, Rhus diversiloba, Rhamnus crocea*). This is the second most common hardwood vegetation state in this zone.
State V. (group [24] & [34]) Woodland, 60-100% canopy closure, few if any shrubs. Understory composed almost exclusively of annual grasses, poison oak is sometimes present but does not exceed 10% cover. This state may occur only on more productive sites in this zone.
State VI. (group [15]) Savanna, 10-20% canopy closure with some understory oaks. Understory dominated by annual grasses, with many shrubs (15-45 % cover) such as *Ceanothus cuneatus, Rhus diversiloba, Rhamnus crocea, Arctostaphylos manzanita*.
State VII. (group [2]) Savanna, 20-60% canopy closure with many understory oaks. Understory dominated by annual grasses and a few shrubs (5-10% cover) such as *Ceanothus cuneatus, Rhus diversiloba, Rhamnus crocea, Arctostaphylos manzanita*.
State VIII. (group [52]) Savannah , 20-60% canopy closure, without understory oaks. Understory dominated by annual grasses, with few shrubs (5-20% cover) such as *Ceanothus cuneatus, Rhus diversiloba, Rhamnus crocea, Arctostaphylos manzanita and Eriodictyon Californicum*.

Catalog of Transitions

Transition 1. (I to II) Colonization of grassland by oak trees.
Although it is known to have occurred over long time periods as vegetation shifted with global climate change, this transition does not occur naturally at the time scale of interest here (< 100 years). In this environment, blue oak seedlings are seldom found further away than 30 m from existing tree canopy.

Planting acorns or seedlings and tending of the young trees using tested methods will achieve this transition on sites that can support trees (time: 30-50 years). A high level of efforts is required for successful implantation of oaks in this low rainfall zone.

Transition 2. (II to I) Complete loss of existing trees without replacement. This transition is very likely since natural regeneration seldom occurs at this level of canopy closure in this precipitation zone. Conditions are also unfavorable for stump resprouting when: trees are *Q. Douglasii* (weak resprouters in general), especially true when drought years follow cutting, trees were large diameter, some specific *Q. Douglasii* genotypes that do not resprout at all are involved.

Loss of trees occur in the following cases:

(a) Type conversion: all trees are cut and/or killed by girdling, herbicides. (transition time: 1 year)

(b) Tree cutting: all trees are cut and resprouting conditions are unfavorable. (time 1-5 years)

(c) Crown fire kills the trees (although likelihood is low given the absence of shrubs and the low canopy closure) and resprouting conditions are unfavorable (time 1-5 years)

(d) Natural mortality: death of the trees over time (time: 50 - 200 years). Tree mortality rate is increased during long periods of drought or following summer rains favoring fungal diseases.

Transition 3. (II to III) Increase in amount of canopy closure by increase of the number of trees

This transition seldom occurs naturally for *Q. Douglasii* at this level of canopy closure. Advance regeneration is very rare and there is no gap effect on recruitment.

However, the transition is possible for *Q. lobata* on sites that are favorable (riparian corridors or deep soils with access to the water table), if a seed source is available and browsing pressure is low (time 30-50 years)

Planting acorns or seedlings and tending of the young trees using tested methods will achieve this transition on sites that can support trees (time: 30-50 years). A high level of efforts is required for successful implantation of oaks in this low rainfall zone.

Transition 4. (III to II) Loss of trees without replacement.

Enough trees are lost in the following cases:

(a) Tree cutting: trees are thinned to below 20% to 30% canopy closure, resprouting conditions are unfavorable and advance regeneration is not present. (transition time 1-5 years)

(b) Crown fire kills the trees (although likelihood is low given the absence of shrubs and the low canopy closure) and resprouting conditions are unfavorable (time 1-5 years)

(c) Natural mortality: death of the trees over time (time: 50 - 200 years). Tree mortality rate is increased during long periods of drought or following summer rains favoring fungal diseases.

Transition 5. (III to IV) Emergence of a cohort of oak saplings in the understory.

The emergence of advance regeneration depends on:

(a) Emergence of seedlings. This occurs when the following conditions are combined: a good acorn year, low levels of acorn predation (few rodents) and above average rainfall the next season. Thinning a stand may increase acorn production.
(b) Browsing pressure is low. This depends on the density of wild browsers (deer in particular) and when livestock are present moderate grazing pressure and removal of grazers once green forage is no longer available decreases browsing (time 2 -10 years).
(c) Protection of seedlings. Can be achieved by a variety of methods.
Transition 6. (IV to III) Disappearance of *Quercus* saplings from understory.
Advance regeneration will vanish from the understory if:
(a) Browsing pressure is high. Under repeated defoliation, oak seedlings and saplings will die or will definitely remain in a shrub form incapable of growing into a tree.
(b) Lack of opening in the canopy (by mortality or thinning) will keep saplings from growing above the browse line.
Transition 7. (IV to V) Increase in amount of canopy closure by increase of the number of trees.
Increase to that level of canopy closure may occur only at particularly productive sites, i.e. with deeper soil, access to a water table, or concentrating run-off from up-slope due to topographic position.
Very low or no browsing.
Transition 8. (V to III) Loss of trees without replacement.
Enough trees are lost when:
(a) Tree cutting: trees are thinned to below 60% canopy closure, resprouting conditions are unfavorable and advance regeneration is not present. (time 1-5 years)
(b) Crown fire kills enough trees (likely at this level of canopy closure, although few shrubs are present to constitute a ladder fuel), resprouting conditions are unfavorable and advance regeneration is not present. (time 1-5 years)
(c) Natural mortality: death of the trees over time (time: 50 - 200 years). Tree mortality rate is increased during long periods of drought or following summer rains favoring fungal diseases.
Transition 9. (VI to VII) Increase of densities of understory *Quercus* saplings and shrubs.
The conditions described for transition 5 will lead to this state over time, low browsing levels or complete protection from herbivory favoring shrubs as well as tree saplings. Low intensity ground fires may favor this transition by reducing competition from annual grasses.
However, overall the density of shrubs is limited by site factors in this zone. The possibility of transition from VII to VIII is unknown at this point.
Transition 10. (VII to III) Disappearance of *Quercus* saplings and shrubs from understory.
Under similar conditions to transition 6. Also, hot burning ground fires may kill many understory trees and shrubs.

Transition 11. (VII to V) Increase in amount of canopy closure by increase of the number of trees. Shrubs cover do not increase and may go down as canopy closes, woodlands with a shrubby understory are rare in this zone.
Increase to that level of canopy closure may occur only at particularly productive sites, i.e. with deeper soil, access to a water table, or concentrating run-off from up-slope due to topographic position.
Transition 12. (VI to VIII) Increase in amount of canopy closure by increase of the number of trees
This transition seldom occurs naturally for *Q. Douglasii* at this level of canopy closure. Although some saplings present in state VI, there is little gap effect on recruitment at this level of canopy closure (transition time 30-50+ years).
Planting acorns or seedlings and tending of the young trees using tested methods will achieve this transition on sites that can support trees (time: 30-50 years). A high level of efforts is required for successful implantation of oaks in this low rainfall zone.
Transition 13. (VIII to VI) Loss of trees without replacement.
See transition 4.
Transition ?. (VI or VIII to I, II, III or VII). Such transitions are doubtful, although not impossible in the experience of the State-and-transition workshop participants. There are some indications that states VI and VIII with their higher levels of shrub cover occur on sites more conducive to shrubs than the majority of sites in this low precipitation zone. Influence of abiotic factors on such sites may more be more alike in effect to that in the next, more mesic zone we have delineated in the Sierra Nevada foothills.

Appendix 4: Data corresponding to density plot in : Figure 5

	31	23	29	5	32	28	21	30	34	22	24	52	4	8	1	2	26	25	40	27	9	42	17	45	38	46	41	44
[28]	46	8	5	13		9	2	2	1	1	3	1	2	1	1	1		1	1	2		0	1					
[32]	20	1	4	9	32	2	1	1	1	1	1	1	2	9	3	3		1	1	3	3	1						
[2]	19			1	4			1		1			1	12	1	7		1	4	1	2		1	1	4	2	1	2
[52	9		1	4	3		2	4		2	1	5	4	8	1	1		5	5	8	7		14	2	1	1	2	1
[33]													6	3				3		6		3	14					3
[15]	13			2	2		1	1		1				10	1	1			1	2			6	3		3		1
[10]				1			1						5	4				1	5		1		11	5	3	2		6
[3]	1			1	1			3					3	14				1	1				1		3			
[20]	5																						35					
[35]																							4		4	8		4
[43]		1											1										8	2	5	8	1	8
[24]	31	6	3	14		3		1	1	7	18		10					4	2		1							
[34]	19	2		10		4			4	2	2	1	23	1			2	7	2	2	4	1	11					
[41]	5	2	2	2									5		2		7		7	14	23	7	11			2	9	2
[42]																				6	17	22	11	6	6			
[36]																	3		3		3	3	3	5	13			
[19]		2											2									2	5	2		11		
[57]				1											1				4		4	1		2	10	6	4	1
[39]	3																						5	5	3	3		
[26]	3	1		4									4				19	12	4		4		12	7	9	1		1
[18]										6								6							6			
[51]		2															6	4	6	2	4	2	2	6	2	4		
[50]		4																					20	2		7	2	
[55]													4										4			4		
[56]																							7		2	3		
[54]		3					3																13			8		

Appendix 4: Data corresponding to density plot in : Figure 5

	10	7	13	14	49	11	53	37	39	48	36	43	19	55	35	57	3	16	54	20	56	47	6	12	15	18	33	50	51	Total
[28]	0																													100
[32]				1																			1				1			100
[2]	1	15	6	4																			1		4					100
[52	6	1																							1					100
[33]	9				3				3											3			11				34			100
[15]	4	4	1	10	1														1			2	5	11	14		2			100
[10]	26	13	3	5	2			1	1														3	2	1					100
[3]	3	3	12	13	3	1		6		1	1			1		10	7	1	1						3					100
[20]	5					5				15								5		30										100
[35]			4	4	8	4				4				4	33			4		8					4					100
[43]	1	3	5	13	3	3	2	1		1		9			1	3		8	3	1	1	1	1	2	1					100
[24]																														100
[34]					1																									100
[41]																														100
[42]					22				6			6																		100
[36]				3	3	5		5		30	13		5				3			3									3	100
[19]			2		5	5		4	2	9		7	16	5	4		4	2		2	2	4	4							100
[57]		1	2		7	6	5	7	6	7	6	1	2		1	12	3				1					1		2		100
[39]		3	5	3	8		8	13	13	8			5	3	10				3									3		100
[26]	4				4		4	1															1							100
[18]						17	6				22		6													33				100
[51]					6	2			2	15	2		4	2												2		2	21	100
[50]	2				4		7		2	7			4							2						2		35		100
[55]	4			33	4						4	4		17				13							8					100
[56]	5	2	13	10			3			2		8			5				2	3	8	7		17	3					100
[54]				5						3						3		15	28	8	10	3		3						100

Appendix4: Tables of data corresponding to density plots in: Figure 9

Strata	Species	[28]	[32]	[2]	[52]	[33]	[15]	[10]	[3]	[20]	[35]	[43]
B	QUDO	83	88	78	58	68	64	67	36	44	17	19
B	QUAG	18	13	19	43	28	17	37	17	59	32	41
B	PISA2	21	32	52	32	28	31	51	31	9	7	19
B	QUKE	4	6	14	18	11	12	22	31		34	43
B	QUWI	10	16	40	23	3	32	33	41	13	12	21
B	QULO	23	11	8	17	14	5	15	5	6	2	8
B	QUCH2		1	6	3	2	4	3	26		10	20
B	PIPO	1	1	7	2	3	3	5	6			14
B	ARME3			1	1	3		2	4	3	5	6
B	UMCA1	0		2		3	0	2	3	6	2	8
B	PSME	2	1	1	3	2	1		3	6	2	5
B	QUDU2	2	5	7	2	3	2		6			3
B	ACMA			1				1	2			
C	QUDO	2	76	62	12	26	42	38	20	9	7	8
C	QUWI	0	13	44	7	3	33	27	40		17	17
C	QUKE		1	6	1	2	6	7	16		34	16
C	QUAG		6	10	5	5	6	6	5	9	27	19
C	QUCH2		1	4	2	2	4	1	20		12	15
C	AECA2		0	8	3	3	9	11	4		12	9
C	UMCA1		2	3		3	3	3	12			8
C	PISA2		6	21	3	6	5	5	4			4
D	RHDI	7	14	31	30	17	31	51	13	44	54	48
D	HEAR2	2	2	13	11	3	29	25	16	22	29	41
D	CECU2	6	8	29	16	12	40	44	19	6	5	20
D	RHCR	5	9	27	12	12	32	28	3	3	15	11
D	RHCA2	2	0	6	3	9	4	14	4	9	22	17
D	ARVI3	1	1	10	8	6	13	27	14	3	29	24
D	ARCA7	3	2	3	21	14	12	19	1	28	20	12
D	CEIN3		0	1	1		3	3	11		17	13
D	ARMA3	3	1	10	8		9	15	5	3	2	7
D	ADFA	1	0	1	1	2	7	5	4	16	5	10
D	HODI			1	1	2	1	5	2	16	5	3
D	BAPI			1	2	3	2	8		22	17	10
D	CHFO2			1			4	4	4		2	12
G	PTAQ			1		10	1	4	1	35	44	4
G	HALI	0	5	3	1	57	6	2		3	15	
G	LOSC	0		1	5	14	2	4	2	22	12	4
G	RUVI2					5	0	1		22	7	3
H	GR2	102	101	101	103	103	103	105	36	22	34	47
Y	LITTER	8	5	41	83	23	36	41	84	53	37	100
Y	BRSOIL	4	2	8	14	12	15	11	41	25	37	52
Y	ROCK	9	14	22	20	12	21	11	13	9	10	18

Appendix4: Tables of data corresponding to density plots in: Figure 9

Strata	Species	[24]	[34]	[41]	[42]	[36]	[19]	[57]	[39]	[26]	[18]	[51]	[50]	[55]	[56]	[54]
B	QUDO	56	69	56	19	14	5	13	11	32	4	12	4	16	27	9
B	QUAG	49	54	47	52	57	61	37	40	71	46	68	74	18	20	35
B	PISA2	22	32	27	11	8	7	12	10	25	7	6	9	16	23	5
B	QUKE	9	8	29	48	44	32	52	63	18	43	26	23	49	35	42
B	QUWI	5	10	20	19	16	12	14	10	8	18	9	5	28	33	6
B	QULO	42	45	29	19	19	3	10	8	34	25	22	16	2	2	2
B	QUCH2	0		4	7	19	17	33	21	5	14	12	9	26	12	6
B	PIPO		2	5	19	11	19	15	24	3	7	10	5	9	8	14
B	ARME3		2	2	15	19	31	15	21	3	14	12	12		6	6
B	UMCA1		1		4	11	17	8	11	7		1	18	2	4	2
B	PSME	2	2	4		5	15	12	6	2	4	3	5	4	1	3
B	QUDU2			4		2	4	3		1				2	1	
B	ACMA		1		11	2		2	3	2	14	22	12	2		
C	QUDO			7			3	8	5	11				12	23	
C	QUWI	0		5			7	16	13	8				46	52	12
C	QUKE						20	20	37	5				33	29	20
C	QUAG			4			37	15	19	15				9	12	9
C	QUCH2			4			12	27	14	2				26	13	11
C	AECA2			2			11	8	17	2				4	21	2
C	UMCA1						17	26	17	9				12	6	2
C	PISA2			2			4	3	2	2				2	9	
D	RHDI		38	36	52	63	55	53	59	49		52	72	30	51	12
D	HEAR2		6	15	11	13	33	20	22	13		14	28	42	41	22
D	CECU2		9	11	15		4	4	16	10		4	7	35	40	23
D	RHCR		14	9	11	10	4	7	14	21		14	16	9	21	6
D	RHCA2		12	13	26	29	20	21	30	23		51	51	16	9	14
D	ARVI3			4	15	10	4	5	21	5		7	7	39	31	31
D	ARCA7		11	15	7	3	9	2	5	11		6	18	4	7	9
D	CEIN3		1		7	5	19	9	24	3		6	9	18	19	23
D	ARMA3		4	2	4	2	1	3	6	2		4	2	5	12	2
D	ADFA		4	2	4	2	1	1	2	3		1	2	19	17	23
D	HODI			2	7	3	9	7	19	2		17	39		5	3
D	BAPI		1	2	11	5	8	4	3	2			5	4	6	11
D	CHFO2		1			2	1	5	22			1	7	9	10	9
G	PTAQ		1		11	22	39		5	6		13	21	7	5	2
G	HALI						1			3						
G	LOSC			2		3	16		2	4			11	2	1	5
G	RUVI2				7	8	35		8	4		9	5		1	3
H	GR2	100	102	102	107	10	31	34	41	104		38	30	33	62	14
Y	LITTER	2	17	85	122	100	100	108	70	75	100	119	86	67	30	37
Y	BRSOIL		2	27	30	17	21	25	21	13	14	4	9	70	24	22
Y	ROCK		8	11	4	11	9	15	10	6		9	4	32	10	6

Appendix 4: Table of data corresponding to the density plot in: Figure 12.

	[28]	[32]	[2]	[52]	[33]	[15]	[10]	[3]	[20]	[35]	[43]	[24]	[34]	[41]	[42]	[36]	[19]	[57]	[39]	[26]	[18]	[51]	[50]	[55]	[56]	[54]
A	39	13	7	5	1	10	3	1			1	9	4	1				1	1	1	1	1		0	2	
B	7	7	10	6	0	16	10	2	0	1	7	1	2	2	1	2	1	3	3	2	1	1	1	3	9	3
C	9		3	12	3	6	52		3	3															9	
D	4	8	8	4		36	3	3	1	1	7	1	1	3		1	1	2		1				6	9	1
E	4	4	11	8	1	11		32			4	2	2	1	3	1		6		3	1	2	2	1		
F				2	2	2	2	4		22	4				2	6	12	4	12			2		16	8	2
G	1	1	2		1	3	1	8		2	15			1		1	3	10	8		1	1	2	8	13	17
H											9					6	19	38	6		6	9	3			3

Appendix 4: Table of data corresponding to the density plot in: Figure 14.

Strata	Species	A	B&D	C	E	F	G	H
B	QULO	8	4	0	13	2	3	3
B	QUDO	86	57	91	78	12	7	0
B	PISA2	22	49	73	61	14	8	6
B	QUWI	25	56	36	86	14	6	0
B	QUKE	6	31	6	17	90	84	97
B	QUCH2	4	15	0	2	10	33	66
B	PIPO	1	12	0	8	29	38	53
B	PSME	1	1	0	4	6	14	31
B	LIDE3	0	2	0	1	0	19	3
C	QUDO	29	34	45	44	8	3	0
C	PISA2	4	7	3	3	0	1	0
C	QUWI	17	49	27	55	12	10	0
C	QUKE	2	13	3	2	69	52	25
C	QUCH2	1	11	0	0	16	33	34
C	AECA2	4	12	12	2	6	4	9
C	PIPO	0	2	0	0	18	16	3
C	LIDE3	0	0	0	0	2	12	0
D	RHDI	14	37	33	20	84	16	50
D	HEAR2	4	29	0	17	12	3	0
D	CECU2	21	42	36	38	10	15	3
D	RHCR	13	19	15	0	0	2	0
D	RHCA2	2	4	6	1	22	7	6
D	ARVI3	3	28	76	35	61	31	13
D	CEIN3	0	8	0	1	45	43	56
D	ARMA3	8	12	36	2	2	3	0
D	CHFO2	0	7	0	0	0	33	6
D	ARPA9	0	1	0	0	6	24	9
D	STOFC	0	0	0	0	41	3	44
D	COCO5	0	1	0	0	12	8	28
D	QUGA2	0	1	0	0	18	6	25
D	RHTR	0	2	3	0	16	1	9
D	CEOC	0	0	0	0	16	2	31
D	CETH	0	0	0	0	18	2	0
G	PTAQP	0	1	0	0	33	8	25
G	BRLA2	0	0	0	0	37	1	0
H	GR2	97	85	94	74	24	28	9
Y	LITTER	16	49	18	99	41	74	100
Y	ROCK	34	21	24	12	18	26	19
Y	BRSOIL	8	18	0	26	29	44	6